Sādhanā Vol. 36, Part 5, October 2011, pp. 651–697. © Indian Academy of Sciences

Epoch-based analysis of speech signals

B YEGNANARAYANA* and SURYAKANTH V GANGASHETTY

International Institute of Information Technology, Hyderabad 500 032, India
e-mail: yegna@iiit.ac.in; svg@iiit.ac.in

Abstract. Speech analysis is traditionally performed using short-time analysis to extract features in time and frequency domains. The window size for the analysis is fixed somewhat arbitrarily, mainly to account for the time varying vocal tract system during production. However, speech in its primary mode of excitation is produced due to impulse-like excitation in each glottal cycle. Anchoring the speech analysis around the glottal closure instants (epochs) yields significant benefits for speech analysis. Epoch-based analysis of speech helps not only to segment the speech signals based on speech production characteristics, but also helps in accurate analysis of speech. It enables extraction of important acoustic-phonetic features such as glottal vibrations, formants, instantaneous fundamental frequency, etc. Epoch sequence is useful to manipulate prosody in speech synthesis applications. Accurate estimation of epochs helps in characterizing voice quality features. Epoch extraction also helps in speech enhancement and multispeaker separation. In this tutorial article, the importance of epochs for speech analysis is discussed, and methods to extract the epoch information are reviewed. Applications of epoch extraction for some speech applications are demonstrated.

Keywords. Epoch; zero-frequency filtering; fundamental frequency; pitch; impulse-like excitation.

1. Significance of epochs in speech analysis

Speech is the output of a time-varying vocal tract system excited by a time-varying excitation. In the resulting speech signal, the information of the speech production system is embedded as relations in the sequence of values of the signal at different instants of sampling the signal. The main objective of speech signal processing is to extract the information of the time varying characteristics of the speech production system. The information is represented in the form of parameters or features derived from the signal. Knowledge at different levels, such as acoustic-phonetic, prosody, lexical, syntactic, etc. is used to interpret the message in the speech signal from the sequence of parameter or feature vectors. Thus, an algorithmic way of extracting the information in the speech signal involves operations of representation (interms of extracted parameters or features) and processing (to extract the information or message), in that order.

*For correspondence

On the other hand, the information or message in the speech, extracted in direct listening by human beings, can be explained in the operations of processing and representation, in that order. This is called as *speech processing* by human beings as opposed to *speech signal processing* by a machine. In speech processing, the input speech acoustic wave is processed first to extract the relevant information such as phonemes, syllables, etc., and the extracted information is represented in the form of sequence of symbols, similar to text as in dictation taking. Thus, speech processing is an evolving human activity, besides the sophisticated auditory processing and selective attention to decide which part is to be processed. It is likely that human processing may not consider all of the acoustic wave input as equally important. The processing may be interpreted in terms of events in the input at various levels, such as at the signal level, speech production level, prosody level, language level, etc. These events may be derived from features present at subsegmental (1–3 ms), segmental (10–30 ms) and suprasegmental (>300 ms) levels.

Methods for speech signal processing mostly use segmental analysis using a fixed frame/window size in the range (10–30 ms). The methods include model-free short-time spectrum analysis or model-based linear prediction (LP) analysis over each (frame size) segment of data (Makhoul 1975). Although the knowledge of speech production and perception is invoked in some analysis techniques such as source-system model or auditory perception based filter-bank type analysis, *all* the speech signal data is used in the initial processing for extracting parameters or features to represent the information in the speech signal. Since human processing of speech seems to be anchored around events, there is need to explore new methods of processing speech signal which are *event-based*.

As mentioned, since events can occur at several levels, it is necessary to first identify such events before further processing of the signal. In this paper, the events at the production level are considered, and in particular, only the significant events occurring due to major source of excitation (glottal vibration) at epochs are considered for further processing to extract information in the speech signal.

Voiced speech analysis consists of determining the glottal pulses representing the excitation source and frequency response of the vocal-tract system. Although, the source of excitation for voiced speech is a sequence of pulses at the glottis, the significant excitation of the vocal-tract system within a glottal pulse can be considered to occur at the instant of glottal closure (GCI), called the *epoch* (Gauffin & Sundberg 1989). Many speech analysis situations depend on accurate estimation of the location of the epoch within a glottal pulse. For example, knowledge of the epoch locations is useful for accurate estimation of the fundamental frequency (F_0). Often the glottal airflow is zero soon after the glottal closure. As a result, the supralaryngeal vocal tract is acoustically decoupled from the trachea. Hence, the speech signal in the closed glottis region represents free resonances of the supralaryngeal vocal-tract system. Analysis of speech signals in the closed glottis regions provides an accurate estimate of the frequency response of the supralaryngeal vocal-tract system (Murty 2009; Veeneman & BeMent 1985). With the knowledge of the epochs, it may be possible to determine the characteristics of the voice source by careful analysis of the signal within a glottal pulse. The epochs can be used as pitch markers for prosody manipulation, which is useful in applications such as text-to-speech synthesis, voice conversion and speech rate conversion (Hamon *et al* 1989; Rao & Yegnanarayana 2006; Yegnanarayana & Veldhuis 1998). Knowledge of the epoch locations may be useful for estimating the time-delay between speech signals collected over a pair of spatially separated microphones (Yegnanarayana *et al* 2005). The segmental signal-to-noise ratio (SNR) of the speech signal is high in the regions around the epochs, and hence it is possible to enhance the speech by exploiting the characteristics of speech signals around the epochs (Yegnanarayana & Murthy 2000). It has been shown that the excitation features derived from the regions around

the epoch locations provide complementary speaker-specific information to the existing spectral features (Murty & Yegnanarayana 2006; Neocleous & Naylor 1998; Plumpe *et al* 1999).

As a result of significant excitation at the epochs, the regions in the speech signal that immediately follow the epochs are relatively more robust to (external) degradations than other regions. The instants of significant excitation play an important role in human perception also. It is because of the epochs in speech that human beings seem to be able to perceive speech even at a distance (e.g., 10 feet or more) from the source, even though the direct signal suffer an attenuation of over 40 dB. For example, we may not be able to get the message in whispered speech by listening to it at a distance of even 10 feet due to absence of regular epochs. The neural mechanism of human beings seems to have the ability of processing selectively the robust regions around the epochs for extracting the acoustic cues even under degraded conditions. It is the ability of human beings to focus on these microlevel events that may be responsible for perceiving speech information even under severe degradation such as noise, reverberation, presence of other speakers and channel variations.

This paper discusses the importance of epochs for speech analysis, reviews epochs extraction methods and illustrates a few applications of epoch-based analysis of speech. The paper is organized as follows (see table 1). Section 2 gives a review of epoch extraction methods which includes methods based on analysis of speech signal. In particular, epoch extraction method based on zero-frequency filtering (ZFF) and the robustness of the method in comparison with other methods are discussed in some detail. Section 3 deals with the main theme of the paper,

Table 1. Table of contents.

namely, how to exploit the epoch location information for speech analysis. It is shown that knowledge of epochs can be used for analysis of speech for glottal activity detection, extraction of instantaneous fundamental frequency, deriving voice quality features from speech such as loudness and characterization of laughter, creaky voice, etc. Some applications are illustrated in section 4, which include time delay estimation, speech enhancement from single and multichannel data and separation of multispeaker data from mixed signals. Section 5 discusses some practical issues in developing speech systems using epoch-based analysis.

2. Review of epoch extraction methods

2.1 *Epoch extraction from electroglottography*

Electroglottography (EGG) is a noninvasive method of measuring the vocal fold contact during voicing without affecting speech production (Abberton *et al* 1989; FFabre 1957; Fourcin & Abberton 1971; Frokjaer-Jensen 1967; Frokjaer-Jensen & Thorvaldsen 1968). The electroglottograph (EGG) measures the variation in impedance to a very small electrical current between a pair of electrodes placed across the neck, as the area of contact of the vocal folds changes during voicing. The demodulated impedance signal is referred to as EGG signal. During voiced speech, the EGG signal exhibits quasi-periodicity according to the frequency of vocal-fold vibration.

Figure 1 shows a few stylized glottal cycles of the EGG signal for a voiced speech segment. The glottal cycle of the EGG signal can be divided into four distinct phases: Closing phase, closed phase corresponding to the region of maximum contact, opening phase and open phase. As long as the glottis is open, the impedance measure across the larynx is maximum, and almost flat (region-4 in figure 1). When the glottis closes, the laryngeal impedance decreases, and the EGG signal shows a steep downward slope (region-1 in figure 1). The opening of the glottis, on the other hand, happens much more gradually (region-3 in figure 1). It should be noted that some authors invert the EGG signal from that shown in figure 1.

According to the theory of voice excitation (Van Den Berg 1958; Stevens 1977), the instant of glottal closure is the point of maximum excitation to the vocal-tract system, and it is used as the starting point of a pitch period. Although the instant of glottal closure is the most abrupt event, it nevertheless needs a finite amount of time. The definition of the starting point of the period, however, requires identification of a unique point in time, that is less subjected to errors.

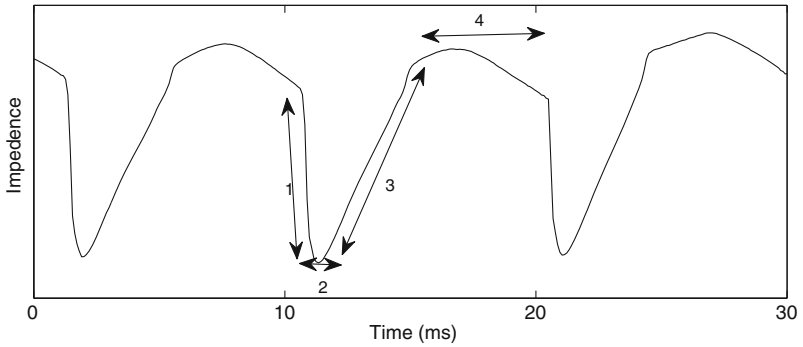

Figure 1. EGG signal for a segment of voiced speech taken from a continuous utterance. Four distinct phases of a glottal cycle in the EGG signal can be identified: closing phase (1), closed phase (2) with maximum contact, opening phase (3) and open phase (4) (Murty 2009, p. 8).

Identifying a unique point directly from the speech waveform is not possible. But such a feature is well manifested in the EGG signal (Scherer *et al* 1988). Since the EGG signal measures directly the laryngeal impedance, it is not affected by the ambient noise. The point of inflection during the steep fall of the EGG signal, i.e., the instant of maximum change of the laryngeal impedance, is typically selected to represent the instant of glottal closure (Lecluse 1977). Hess & Indefrey (1987) defined an epoch to occur at the maximum of the time-derivative of the smoothed EGG signal during a glottal cycle. Huckvale (2000) developed an algorithm that identifies epoch locations as the positive-going zero-crossings in the smoothed time-derivative of the EGG signal.

Figure 2 shows a segment of voiced speech, its EGG signal and the differenced EGG signal. The locations of sharp negative peaks in the differenced EGG signal denote the instants of glottal closure. The negative peak amplitude of the differenced EGG signal denotes the maximum flow declination rate, which can be hypothesized to be the strength of the epoch. It can be noticed that, in contrast to the speech signal, the EGG signal is hardly affected by the time-varying vocal-tract system. Also changes in the shape and amplitude of the cycles in EGG are relatively small. Hence, the epoch locations and their strengths can be accurately determined from the EGG signal even in the dynamic regions where the vocal-tract system is not stationary.

Since every glottal cycle is represented by a single pulse, the EGG signal can be used for accurate determination of instantaneous fundamental frequency of the voiced speech segments. In addition, the EGG signal provides the basis for a good voiced–unvoiced discrimination, since the differenced EGG signal is almost zero during unvoiced segments, where the glottis is always open. Since the EGG signal is not normally available in practice, there exists strong motivation

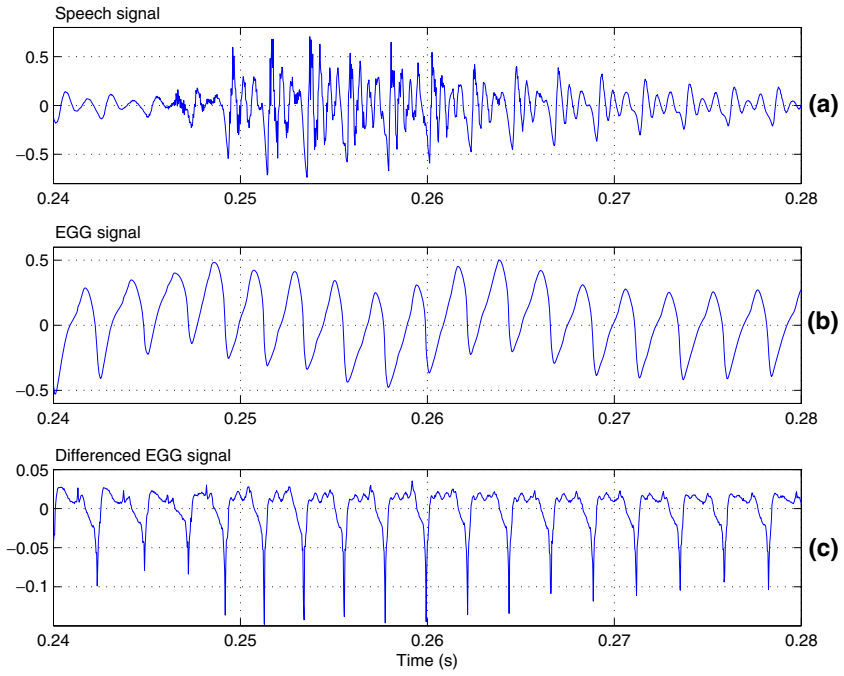

Figure 2. Extraction of epoch locations from differenced EGG signal. (**a**) A segment of voiced speech taken from a continuous utterance, (**b**) EGG signal, and (**c**) differenced EGG signal. Locations of the negative peaks in the differenced EGG signal correspond to the instants of glottal closure (Murty 2009, p. 9).

to develop techniques for extracting these source features from the speech signal. Some of the techniques presented in the literature are reviewed below.

2.2 *Epoch extraction from speech signals*

Several methods have been proposed for estimating the instants of glottal closure from speech signal without using the EGG signal. For convenience, these methods are categorized as follows: (i) methods based on short-time energy of the speech signal, (ii) methods based on predictability of an all-pole linear predictor, and (iii) methods based on properties of group-delay. It should be noted that the methods placed in one category could also belong to another, depending on the interpretation of the method.

2.2a *Epoch extraction from short-time energy of speech signal*: Glottal closure instants can be detected from the energy peaks in waveforms derived directly from the speech signal (Jankowski Jr *et al* 1995; Ma & Willems 1994) or from the features in its time–frequency representation (Navarro-Mesa *et al* 2001; Tuan & d'Alessandro 1999). In Ananthapadmanabha & Yegnanarayana (1975), a method based on the composite signal decomposition was proposed for epoch extraction of voiced speech. A superposition of nearly identical waveforms was referred to as a composite signal. The epoch filter proposed in this work, computes the HE of the highpass filtered composite signal to locate the epoch instants. It was shown that the instants of excitation of the vocal tract could be identified precisely even in continuous speech. However, this method is suitable for analysing only clean speech signals.

The Frobenius norm offers an estimate of the short-term energy of the speech signal. The Frobenius norm computed using a sliding window gives an estimate of the energy value at every speech sample. The locations of the peaks in the energy signal indicate the glottal closure instants. Frobenius norm approach based on singular value decomposition was proposed in Ma & Willems (1994). The method was shown to work only for vowel segments. No attempt was made to detect epochs for difficult cases such as nasals, voiced consonants and semivowels, as in these cases, due to loading of the glottis by the vocal tract system, the strength of excitation at the epoch is not strong as in the case of vowels.

The energy peaks can also be detected in a time–frequency representation of the speech signal. Wavelet transform has been used to represent the speech and to detect the glottal closure instants (Tuan & d'Alessandro 1999). Lines of amplitude maxima in the time–frequency plane were identified, and the epochs were determined using the line carrying the maximum accumulated amplitude within each pitch period. Alternatively, a Cohen's class time–frequency representation of speech was constructed and used to detect the epochs (Navarro-Mesa *et al* 2001). The epochs were detected as peaks in a spectral density correlator derived from the time–frequency representation.

2.2b *Epoch extraction from LP*: Many methods for epoch extraction rely on the discontinuities in a linear model of speech production. An early approach used a predictability measure to detect epochs by finding the maximum of the determinant of the autocovariance matrix of the speech signal (Sobakin 1972; Strube 1974).

The error signal obtained in the LP analysis, referred to as the LP residual, is known to contain information pertaining to epochs. A large value of the LP residual within a pitch period is supposed to indicate the epoch location (Atal & Hanauer 1971). However, epoch identification directly from the LP residual is not recommended (Strube 1974), because the LP residual contains peaks of random polarity around the epochs. Figure 3b shows the LP residual derived

through a 10^{th}-order LP analysis of the speech segment shown in figure 3a. The epoch locations cannot be unambiguously identified from the LP residual shown in figure 3b, because of the occurrence of multiple peaks of either polarity. The reference epoch locations shown by the differenced EGG signal is given in figure 3d.

A method for unambiguous identification of epochs was proposed in Ananthapadmanabha & Yegnanarayana (1979) by computing the amplitude envelope of the analytic signal of the LP residual, referred to as the Hilbert envelope (HE) of the LP residual. Figure 3c shows the HE of the LP residual signal shown in figure 3b. Unlike the LP residual, the HE shows sharp unambiguous peaks which are in close agreement with the reference epoch locations, as shown by the differenced EGG signal in figure 3d. Although this method works well on clean signals, performance of the method degrades under noisy conditions due to sensitivity of the LP analysis to noise in the signal.

Covariance analysis in the least squares approach for accurately performing the glottal inverse filtering of the acoustic speech waveform is used in Wong *et al* (1979). In this work, epochs were detected based on a measure derived from the total energy of the LP residual derived over a sliding window. This method was further enhanced in Plumpe *et al* (1999), using the observation that the formant modulations are slower in the closed phase region than in the open phase region (Ananthapadmanabha & Fant 1982).

One of the difficulties in using the prediction error for epoch extraction is that it often contains effects due to resonances of the vocal-tract system, as the derived inverse filter does not completely suppress the formant frequencies. As a result, the excitation peaks become less prominent in the residual signal. In an attempt to overcome this limitation, a method based on maximum likelihood theory for epoch determination is proposed in Cheng & O'Shaughnessy (1989). In this method, the speech signal was processed to get the maximum-likelihood epoch detection (MLED) signal. The strongest positive pulse in the MLED signal indicates the epoch location within a pitch period. The MLED signal creates not only a strong and sharp pulse at epoch,

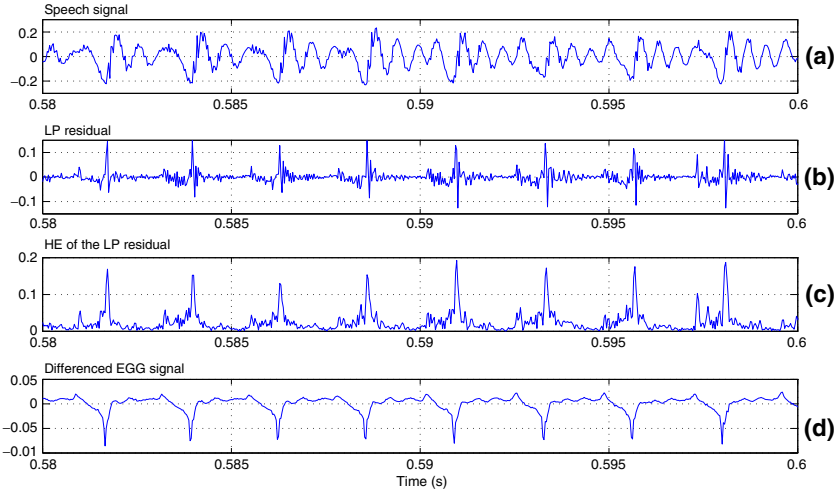

Figure 3. HE of LP residual of the speech signal for epoch extraction. (**a**) A segment of voiced speech signal taken from a continuous utterance. (**b**) LP residual. (**c**) HE of the LP residual. (**d**) Differenced EGG signal for observing the reference epoch locations.

but also a set of weaker pulses that represent suboptimal candidates for epoch within a pitch period. Hence, a selection function was derived using the speech signal and its Hilbert transform, that emphasized the contrast between the epoch and the suboptimal pulses. The limitation of this method is the choice of the window for deriving the selection function, and also the use of threshold for deciding the epochs.

Kalman filtering has been applied to detect the closed phase regions in voiced speech (McKenna 2001). The boundaries of the closed phase, i.e., the instant of glottal closure and the instant of glottal opening, are detected using the logarithm of the determinant of the error covariance matrix of the Kalman filter. This measure assesses the predictability of the speech signal, and is able to detect the glottal closure instants, but the timing accuracy is poor.

2.2c *Epoch extraction using group-delay functions*: A method for detecting the epochs in a speech signal using the properties of minimum phase signals and group-delay function was proposed in Smits & Yegnanarayana (1995). The method is based on the fact that the average value of the group-delay function of a signal within an analysis frame corresponds to the location of the significant excitation. An improved method based on the computation of the group-delay function directly from the speech signal was proposed in Yegnanarayana & Smits (1995). The average value of the group-delay function is called the phase slope value, and it is computed for each sample shift to obtain the phase slope function. The instants of zero crossings of the phase slope function correspond to epochs. Robustness of the group-delay based method against additive noise and channel distortions was studied in Murty & Yegnanarayana (1999). Four measures of group-delay (average group-delay, zero-frequency group-delay, energy weighted group-delay and energy weighted phase) and their use for epoch detection were investigated in Brookes *et al* (2006). The effect of the length of the analysis window, the trade-off between the detection rate and the timing error, and the computational cost of evaluating the measures were also examined in detail. It was shown that the energy weighted measures performed better than the other two measures.

2.3 *DYPSA algorithm for epoch extraction*

A dynamic programming projected phase-slope algorithm (DYPSA) for automatic estimation of glottal closure instants in voiced speech was presented in Kounoudes *et al* (2002) and Naylor *et al* (2007). The candidates for GCI were obtained from the zero-crossings of the phase-slope function derived from the energy weighted group-delay, and were refined by employing a dynamic programming algorithm. It was shown that DYPSA performed better than the existing methods.

Epoch is an instant property. However, in most of the methods discussed above (except the group-delay-based methods), the epochs are detected by employing block processing of signal or its LP residual, which may result in ambiguity about the precise locations of the epochs. Moreover, the excitation impulses need not be periodic, although some of these methods exploit periodicity for accurate estimation of the epoch locations. In general, it is difficult to detect the epochs in the case of low-voiced consonants, nasals and semivowels, breathy voices and female speakers. A new method which overcomes some of these limitations is discussed in the next section.

2.4 *Epoch extraction using zero-frequency resonator*

The discontinuities in the excitation signal caused by the sharp closure of the glottis can be approximated by a sequence of impulses of varying amplitudes. The effect of the impulse-like

excitation is reflected across all frequencies, including the zero-frequency (0 Hz). The effect of the discontinuity due to the impulse-like excitation is clearly visible in the output of narrowband filtering (i.e., near ideal resonator at any frequency) of the speech signal. The advantage of choosing the zero-frequency (0 Hz) resonator is that the characteristics of the time-varying vocal-tract system will not affect the characteristics of the discontinuities in the output of the resonator. This is because the vocal-tract system has resonances at much higher frequencies than at the zero-frequency. An ideal zero-frequency digital resonator is an infinite impulse response filter with a pair of poles located on the unit circle. The decay in the frequency response is determined by the digital approximation of the all-pole filter. A cascade of two such resonators is used to provide sharper cut-off to reduce the effect of resonances of the vocal-tract system. The following steps are involved in processing speech signal to derive the zero-frequency filtered signal (Murty & Yegnanarayana 2008; Yegnanarayana *et al* 2008).

(i) The speech signal $s[n]$ is differenced to remove any slowly varying component introduced by the recording device.

$$x[n] = s[n] - s[n-1] \tag{1}$$

(ii) The differenced speech signal $x[n]$ is passed through a cascade of two ideal zero-frequency (digital) resonators, i.e.,

$$y_0[n] = -\sum_{k=1}^{4} a_k y_0[n-k] + x[n], \tag{2}$$

where $a_1 = -4$, $a_2 = 6$, $a_3 = -4$ and $a_4 = 1$. The resulting signal $y_0[n]$ grows approximately as a polynomial function of time.

(iii) The average pitch period is computed using the autocorrelation function of 30 ms segments of $x[n]$.

(iv) The trend in $y_0[n]$ is removed by subtracting the local mean computed over the average pitch period, at each sample. The resulting signal

$$y[n] = y_0[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} y_0[n+m] \tag{3}$$

is the zero-frequency-filtered (ZFF) signal. Here, $2N+1$ corresponds to the number of samples in the window used for trend removal. The choice of the window size is not critical as long as it is in range of one to two pitch periods. Figure 4b shows the ZFF signal of the speech segment shown in figure 4a. It was shown that the instants of positive-to-negative zero crossings (PNZCs) correspond to the instants of significant excitation in voiced speech, or *epochs* (Murty & Yegnanarayana 2008). The locations of the PNZCs in the ZFF signal are shown by downword arrows in figure 4c. There is close agreement between the locations of the strong positive peaks of the negative of the differenced electroglottograph (DEGG) signal and the instants of PNZCs derived from the ZFF signal. Information about the strength of impulse-like excitation can also be derived from the ZFF signal, using the slope of the ZFF signal around PNZCs (Murty *et al* 2009). The slope is measured by computing the difference between the negative and positive sample values on either side of the epoch, and is denoted as strength of impulse (ϵ) (Murty *et al* 2009). Figure 4d shows the plot of ϵ, derived from the ZFF signal in figure 4b. The plot of ϵ shows a trend similar to the DEGG signal (figure 4c).

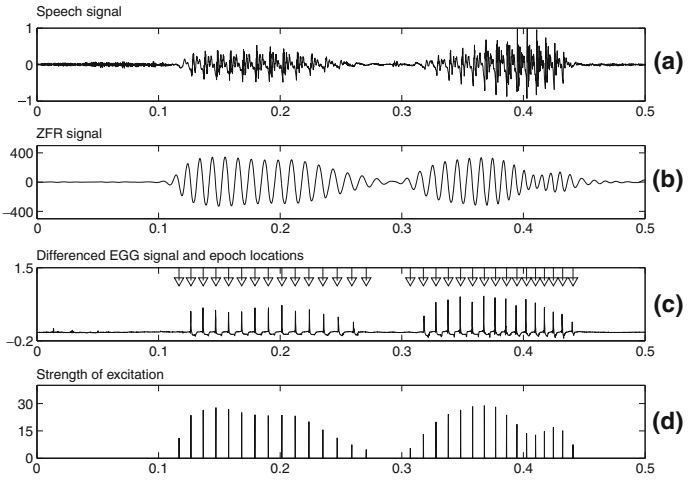

Figure 4. (**a**) Segment of a speech signal from VOQUAL'03 (d'Alessandro & Scherer 2003) database, (**b**) zero-frequency filtered signal, (**c**) negative of the differenced EGG signal and epoch locations marked by arrows, and (**d**) strength of excitation of impulse calculated from the zero-frequency filtered signal.

It is interesting to note that even for an aperiodic sequence of impulse-like excitations, the positive zero-crossings of the ZFF signal correspond to the locations of the epochs (Murty 2009). There is no such relation between the excitation and the ZFF signal for random noise excitation of the time-varying all-pole system. Further, the ZFF signal has significantly lower amplitudes for the random noise excitation compared to the impulse sequence excitation (Murty 2009). This is the case for unvoiced segments of speech.

2.5 *Comparison of epoch extraction methods*

In this section, the zero-frequency-based epoch extraction method is compared with three existing methods in terms of identification accuracy and robustness against degradation. The three methods chosen for comparison are the HE-based method (Rao *et al* 2007), the group-delay (GD)-based method (Smits & Yegnanarayana 1995) and the DYPSA algorithm (Naylor *et al* 2007).

The CMU-Arctic database (Kominek & Black 2004) is used to evaluate the proposed method of epoch extraction, and to compare the results with the existing methods. In addition to the

Table 2. Performance comparison of epoch extraction methods on CMU-Arctic database. IDR – Identification rate, MR – Miss rate, and FAR – False alarm rate, and IDA – Identification accuracy (Murty & Yegnanarayana 2008).

Method	IDR (%)	MR (%)	FAR (%)	IDA (ms)
HE-based	89.86	1.43	8.71	0.58
GD-based	92.80	4.01	3.18	0.67
DYPSA	96.66	1.76	1.58	0.59
Zero-frequency-based	99.04	0.18	0.77	0.36

Table 3. Performance comparison for epoch detection methods for various SNRs and noise environments. IDR – Identification rate (from Murty 2009, p. 67).

Environment		IDR (%)			
Noise	SNR (dB)	HE-based	GD-based	DYPSA	Zero-frequency method
White	20	84.56	87.34	92.12	99.04
White	15	82.26	84.65	85.33	99.06
White	10	79.45	81.07	75.95	99.05
Babble	20	86.73	89.45	96.42	99.02
Babble	15	84.88	87.27	96.14	98.99
Babble	10	82.51	84.32	95.48	98.83
HF Channel	20	84.23	86.54	95.89	99.04
HF Channel	15	82.04	83.87	94.99	99.05
HF Channel	10	79.24	80.13	92.4	99.06
Vehicle	20	89.75	92.67	96.67	99.06
Vehicle	15	89.58	92.49	96.6	98.93
Vehicle	10	89.25	92.18	96.64	97.83

speech signals, the Arctic database contains simultaneous recordings of EGG signals. Reference locations of the epochs were extracted from the voiced segments of the EGG signals by finding peaks in the negative of the differenced EGG signal. The performance of the algorithms was evaluated only in the voiced segments (detected from EGG signal) between the reference epoch locations and the estimated epoch locations. The database contains a total of 792249 epochs in the voiced regions.

Table 2 shows the performance results on Arctic database for identification rate, miss rate, false alarm rate and identification accuracy for the three existing methods, HE-based, GD-based and DYPSA algorithm, as well as for the zero-frequency method. From table 2, the zero-frequency method of epoch extraction gives better identification rate as well as identification accuracy, compared to all the three methods.

The effect of noise on the accuracy of epoch detection methods is evaluated using artificially generated noisy speech data. Several noise environments at varying SNR were simulated to evaluate the robustness of the epoch detection methods. Table 3 shows the comparative performance of epoch extraction methods for different types of degradations at varying SNRs. The zero-frequency method consistently performs better than the existing techniques even under degradation.

3. Epoch-based speech analysis

In this section, speech analysis methods are discussed to illustrate the significance of epochs.

3.1 *Glottal activity detection*

The primary mode of excitation in speech production is due to vibration of vocal folds at the glottis, and hence may be considered as glottal activity. The strength of impulses during the glottal activity is determined mostly by the rate of closure of the vocal folds in each glottal cycle.

In the absence of vocal fold vibration, the vocal-tract system may be excited by random noise, as in the case of frication. The energy of the random noise excitation is distributed both in time and frequency domains. Whereas the energy of an impulse is highly concentrated in the time domain, but is distributed uniformly in the frequency domain. The ZFF signal exhibits significantly lower amplitudes for random noise excitation compared to the impulse-like excitation, and hence can be used to detect the regions of glottal activity (vocal fold vibration) as illustrated in figure 5. Figure 5a shows a segment of speech signal with regions of glottal activity, marked by dotted lines, obtained from the differenced EGG signal in figure 5b. The ZFF signal of speech, shown in figure 5c, clearly indicates the regions of glottal activity, and the regions match well with those obtained from the differenced EGG signal in figure 5b. It can be noticed that the unvoiced regions around 0.6 s and 1.2 s in the speech signal (figure 5a) have very low amplitude in the ZFF signal (figure 5c). Hence, the short-term energy of the ZFF signal computed over 20 ms frames, shown in figure 5d, can be used for glottal activity detection (GAD).

The above GAD method was evaluated under different noisy environments (Varga & Steeneken 1993) at varying levels of degradation. A subset of CMU-Arctic database (Kominek & Black 2004) consisting of 100 randomly selected sentences from each of the three speakers was used to evaluate the GAD method. The entire dataset was sample-wise labelled for glottal activity using the simultaneously recorded EGG signals available with the database.

The performance of the proposed GAD method was evaluated using the detection error trade-off (DET) curves (Martin *et al* 1997), which show the trade-off between false alarm rate (FAR) and false rejection rate (FRR). The performance of the system is expressed in terms of equal error rate (EER), the point at which FAR and FRR are equal. The lower the EER value, the higher is the accuracy of the GAD method. Figure 6 shows the DET curves obtained for the GAD algorithm under different noise environments at an SNR of 0 dB. The performance of GAD at varying levels of degradation is listed in table 4 using the reference derived from the EGG signals.

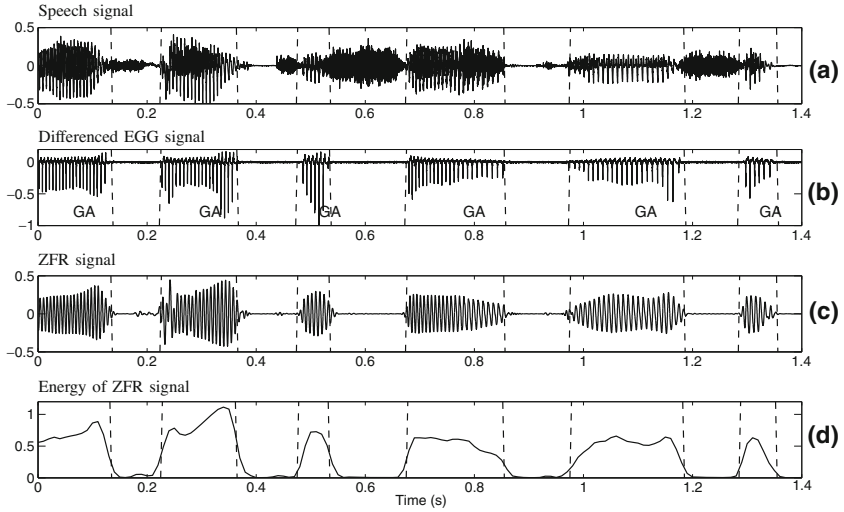

Figure 5. Glottal activity detection from the filtered signal. (**a**) Speech signal. (**b**) Differenced EGG signal. (**c**) Zero-frequency filtered signal. (**d**) Energy computed over 20 ms segments of the filtered signal. Regions marked as GA in (**b**) indicate regions of glottal activity (Murty 2009).

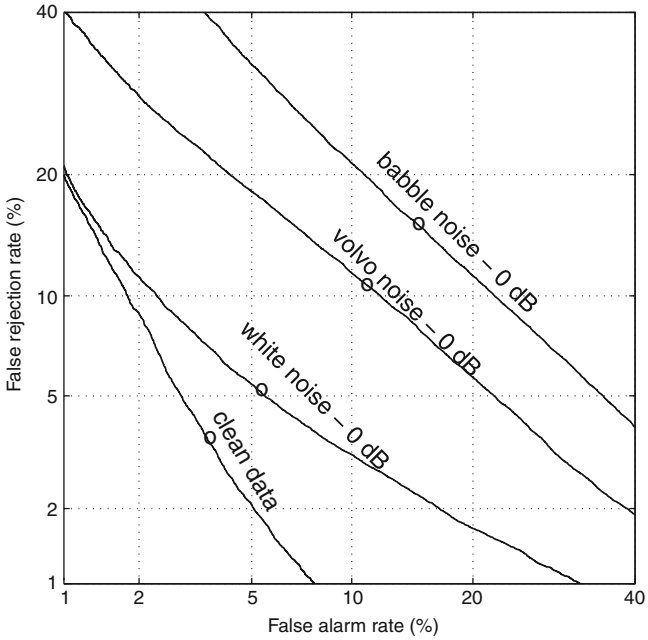

Figure 6. DET curves indicating the performance of GAD method under different noise environments (Murty 2009). Circles in figure correspond to EERs.

3.2 *Estimation of instantaneous fundamental frequency*

The duration of each glottal cycle is the fundamental period, and the reciprocal of this period is referred to as the fundamental frequency. Fundamental frequency estimation is often associated with voicing decision to eliminate the unvoiced regions from considerations. The glottal activity detection discussed in section 3.1 is used to detect the voiced regions. Since the glottal activity detection is based on the strength of the impulse in each glottal cycle, the end-points of the voiced regions can be obtained accurately.

To reduce the effects due to spurious zero-crossings in the filtered signal, the HE of speech signal is also used. The HE contains a sequence of strong impulses around the glottal closure instants, and may also contain some spurious impulses at other places due to the formant structure of the vocal tract, and the secondary excitations in the glottal cycles. However, the amplitudes of the impulses around the glottal closure instants dominate over those of the spurious impulses in the computation of the ZFF signal. Hence, the ZFF signal of the HE mainly contains

Table 4. Performance of GAD in EER (%) under different noise environments at varying levels of degradation. Reference is derived from EGG signals (Murty 2009).

Noise type	20 dB	15 dB	10 dB	5 dB	0 dB
White	3.56	3.56	3.60	3.78	5.24
Babble	3.56	3.64	4.62	7.95	15.10
Vehicle	3.56	3.58	4.09	6.28	10.83

the zero-crossings around the instants of glottal closure. However, the zero-crossings derived from the ZFF signal of HE are not as accurate as those derived from the ZFF signal of speech signal. Hence, the accuracy of the zero-crossings derived from the filtered signal of speech, and the robustness of the zero-crossings derived from the HE are used in conjunction to obtain an accurate and robust estimate of the instantaneous fundamental frequency (Yegnanarayana & Murty 2009).

The instantaneous pitch frequency contour obtained from the ZFF signal of speech is used as the primary pitch contour, and the errors in the contour are corrected using the pitch contour derived from the HE of the speech signal. The pitch frequency contours are obtained for every 10 ms from the zero-crossings of the ZFF signals. The value of 10 ms is chosen for comparison with the results from other methods. Let $p_s[m]$ and $p_h[m]$ be the pitch frequency contours derived, respectively, from the speech signal and the HE of the speech signal. The following logic is used to correct the errors in $p_s[m]$:

$$p[m] = \begin{cases} p_h[m], & \text{if } p_s[m] > 1.5\,p_h[m] \\ p_s[m], & \text{otherwise,} \end{cases} \tag{4}$$

where m is the frame index for every 10 ms, and $p[m]$ is the corrected pitch contour. The factor 1.5 is used mainly to reduce the pitch doubling errors in $p_s[m]$ due to spurious zero-crossings. Any value between 1.3 and 1.8 is adequate to perform this correction.

Results of the above epoch-based method of extracting the instantaneous F_0 are compared with the results from four other methods available in the literature, in terms of accuracy in estimation and robustness against degradation. The four methods are: (i) Praat's autocorrelation method (Boersma 1993), (ii) crosscorrelation method (Goldberg & Riek 2000), (iii) subharmonic summation (SHS) (Hermes 1988), and (iv) a fundamental frequency estimator (YIN) (de Cheveigne & Kawahara 2002). The performance of the epoch-based method and the other four methods is evaluated on two sets of databases: Keele database (Plante *et al* 1995) and the CSTR database (Bagshaw *et al* 1993). Table 5 gives the performance of pitch contours derived from the five methods. The percentage gross errors for the epoch-based method are significantly lower than the percentage gross errors for other methods.

The effect of noise on the accuracy of the pitch estimation algorithms is examined for three different types of noises (white Gaussian noise, babble noise and vehicle noise) collected from

Table 5. Performance of algorithms for fundamental frequency estimation on clean data. $p_s[m]$ denotes the pitch contour derived from filtered speech signal. $p_h[m]$ denotes the pitch contour derived from filtered HE. $p[m]$ denotes the pitch contour obtained by combining evidence from $p_s[m]$ and $p_h[m]$ as in Eq. 4 (Murty 2009).

	Keele Database			CSTR Database		
Method	GE (%)	M (Hz)	SD (Hz)	GE (%)	M (Hz)	SD (Hz)
AC	5.345	2.656	3.694	5.238	4.777	6.820
CC	6.891	2.201	3.371	6.818	5.108	6.730
YIN	3.219	2.165	2.906	3.073	4.922	6.584
SHS	10.774	1.868	2.398	8.938	4.108	5.864
$p_s[m]$	2.935	3.198	4.555	3.394	5.459	6.974
$p_h[m]$	5.647	4.562	6.381	4.157	5.699	6.886
$p[m]$	2.603	3.207	4.473	1.943	5.367	6.801

Noisex-92 database (Varga & Steeneken 1993). The results for different SNRs given in Murty (2009) establish the robustness of the epoch-based method which can be attributed to the impulse-like nature of the glottal closure instants in the speech signal, as all the other methods depend mostly on the periodicity of the signal in successive glottal cycles. It should be noted that the periodicity of the signal waveform is affected by noise.

3.3 *Estimation of formant frequencies*

Owing to time varying vocal-tract system during production of speech, the resonances (formants) of the vocal tract change continuously with time. In fact, even within a glottal cycle there is significant change in the size and shape of the vocal tract in the closed and open phase of the glottis. The epoch locations help in locating these regions, as the region immediately after the epoch corresponds mostly to the closed phase of the glottis, where the formants correspond to the size and shape of the vocal tract above the glottis. These formants are also stronger and sharper (lower bandwidth) compared to the formants in the open glottis region, which is the region before the epoch. Although epoch locations help to separate the open and closed glottis regions, it is a challenge to extract the formant frequencies from short (<3 ms) segments of the closed glottis region of the speech signal.

The characteristics of the phase of Fourier transform can be exploited for extracting the formant frequencies from short segments of speech (Yegnanarayana 1978). The problem of phase wrapping can be overcome by using group delay function, which is defined as the negative derivative of the phase of the Fourier transform of the signal (Joseph *et al* 2006). The group delay function $\tau(\omega)$ for a resonator with magnitude response $|H(\omega)|$ is directly proportional to the square of the magnitude response around the resonance frequency ω_0, i.e., $\tau(\omega_0) \propto |H(\omega_0)|^2$ (Yegnanarayana 1978). Further, the group delay function for a cascade of resonators is the sum of the group delays of the individual resonators. The high resolution and the additive properties of the group delay function are exploited to derive the formant frequencies directly from the speech signal. The group delay function $\tau(\omega)$ is computed from the speech signal $x[n]$ using the following relation (Guruprasad 2010):

$$\tau(\omega) = \frac{X_I(\omega)Y_R(\omega) - X_R(\omega)Y_I(\omega)}{X_R^2(\omega) + X_I^2(\omega)}. \tag{5}$$

Here, $X(\omega) = X_R(\omega) + jX_I(\omega)$ is the Fourier transform of $x[n]$, while $Y(\omega) = Y_R(\omega) + jY_I(\omega)$ is the Fourier transform of $y[n] = -jnx[n]$. The computation of the group delay function is affected by the zeros in the magnitude spectrum ($|X^2(\omega)| = X_R^2(\omega) + X_I^2(\omega)$). On the other hand, if we consider only the numerator term of the group delay function, i.e., $g(\omega) = X_I(\omega)Y_R(\omega) - X_R(\omega)Y_I(\omega)$, then $g(\omega) \propto |X(\omega)|^4$, since $\tau(\omega) \propto |X(\omega)|^2$ near the resonance frequency (Joseph *et al* 2006). This indicates that the numerator of the group delay function has higher resolution than the group delay function itself. This property of the group delay function can be exploited to derive formant frequencies from short segments of speech signal at every epoch. Figure 7 shows the extracted formants from the peaks of $g(\omega)$ computed at each epoch. It should be noted that, here the formant contours are derived in a nonparametric way, as no model (as in linear prediction) is assumed to extract the formants.

Since the regions around the epochs also correspond to high SNR regions within each glottal cycle, it is also possible to extract the formants even from degraded speech, such as speech collected at a distance (say 6 feet) from the speaker, even though the distant speech signal is degraded by noise and mild reverberation. This is possible, provided the epochs are extracted

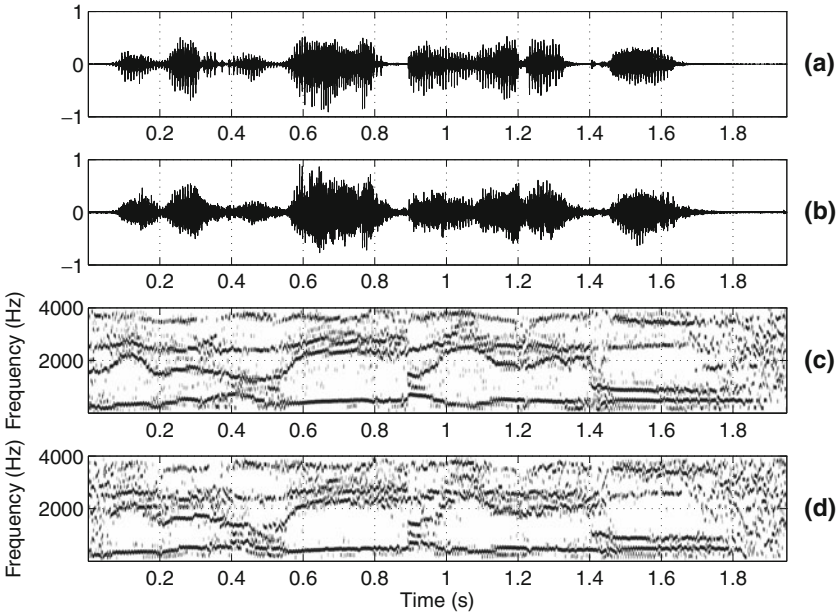

Figure 7. (**a**) A segment of close-speaking speech signal and (**b**) the corresponding distant-speaking speech signal. Contours of formants derived from (**c**) the close-speaking speech signal and (**d**) the distant speech signal (Guruprasad & Yegnanarayana 2010).

from distant speech (Guruprasad & Yegnanarayana 2010). Figure 7 shows the formant contours from distant (6 feet) speech signal. It is observed that the formant contours are similar to those obtained from the close speaking speech signal.

3.4 *Glottal activity in stop consonant analysis*

The primary and most important mode of excitation is due to the activity at the glottis. In normal voiced excitation (called modal voicing), there will be vibrations of the vocal folds resulting in glottal opening, followed normally by an abrupt closure of the vocal folds, and then a closing phase of the glottis, before the glottis is opened again for the next cycle due to build-up of pressure from the lungs. Other aspects of glottal activity include vibration with large opening for production of breathy voice, a complete opening for the production of unvoiced sounds, a partial closure of the vocal folds for production of creaky voice, and finally a complete closure of the vocal folds such as for glottal stops. Figure 8 illustrates the continuum of phonation types as proposed by Gordon & Ladefoged (2001). In this subsection, the focus is on extraction of the excitation characteristics due to glottal activity, in order to derive the acoustic correlates of stop consonants using the excitation information.

Figure 8. Phonation types (Gordon & Ladefoged 2001).

Stop consonants are a class of speech sounds whose characteristic feature is an interval during which the airflow is completely blocked within the oral cavity. The air pressure built up behind the oral closure is released more or less impulsively as the vocal tract moves towards a configuration appropriate for the following vowel. Depending on the place of closure in the oral cavity, different linguistic contrasts of the stop consonants can be produced. A stop consonant is said to be voiced (in a phonetic sense) if there is an audible laryngeal pulsation during the closure phase, and unvoiced if it is absent. Another phonetic feature traditionally attributed to the stop consonants is aspiration.

Figure 9 illustrates the relative placement of the important events in the production of various stop consonants. The interval between the time of burst release to the time of onset of vocal fold vibration is called the voice onset time (VOT) (Ambramson & Lisker 1965). The VOT is a commonly used feature to analyse the manner of articulation in the production of stop consonants.

The two modes of excitation in the production of stop consonants are: (i) vocal fold vibration, and (ii) burst and aspiration. The regions of vocal-fold vibration can be extracted from the zero-frequency filtered signal. The burst and aspiration regions correspond to band-limited noise, and can be analysed using the LP residual. The normalized error $\eta(n)$ is defined as the ratio of the energy of the LP residual and the speech signal for a block or frame of data of size 20 ms centred around the sample at the instant n. The normalized error is computed for every sample shift of the frame. Figure 10 illustrates the excitation source features for the four velar stop consonants /ka/, /k^{h}a/, /ga/, /g^{h}a/ in Indian languages. The plots of the filtered output clearly shows the regions of glottal activity.

The following observations can be made to distinguish the four categories:

(i) *Unvoiced unaspirated:* There is sudden increase in the normalized error at the release of the burst. The normalized error is large in the short burst region relative to the modal voicing region.

(ii) *Unvoiced aspirated:* There is sudden increase in the normalized error at the release of the burst. The large $\eta[n]$ is extended over the aspirated region due to the presence of breathy

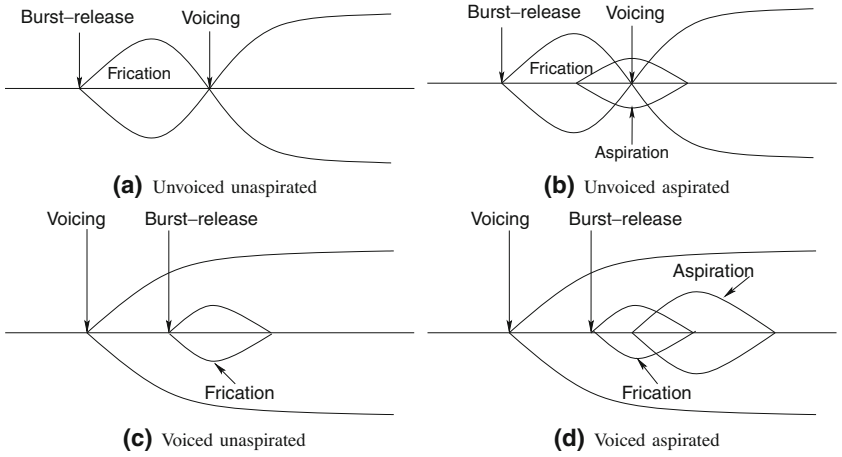

Figure 9. Schematic representation of the important events in the stop consonants (Murty 2009).

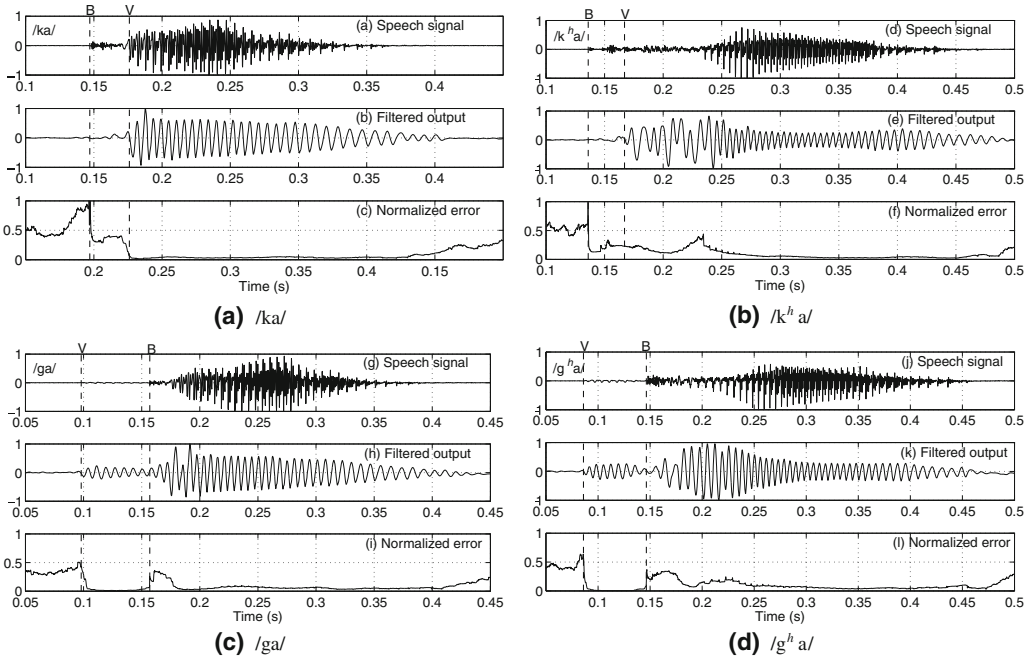

Figure 10. The speech signal, filtered (ZFF) output, and the normalized error for four different velar stop sound units (Murty 2009).

noise. The $\eta[n]$ is low in the modal voicing region. The ZFF output is *somewhat less periodic* in the aspirated region.

(iii) *Voiced unaspirated:* There is relatively large output in the ZFF signal due to initial voicing compared to the relatively small amplitude in the waveform. There is increase in the $\eta[n]$ during the short burst region.

(iv) *Voiced aspirated:* The ZFF output is large during the initial voicing region, and then in the aspirated and modal voicing regions. There is a dip in the ZFF output at the burst release. However, the $\eta[n]$ has an abrupt raise at the burst release, followed by large $\eta[n]$ in the aspirated region due to breathy noise.

Similar features are observed in stop consonants produced at other places of articulation. The burst release instant (marked as B) is determined as the instant where there is a large increase in $\eta[n]$. The starting instant of the glottal activity (marked as V) is derived from the ZFF output. The interval between these two instants can be used as VOT to characterize the different stop consonants.

3.5 *Determination of phonetic features of glottalization, creaky voice and glottal stop*

For producing normal pulmonic stop consonants (discussed in section 3.4), the articulators form an oral closure, building up pressure behind the closure by the air coming from lungs, and the built-up pressure is released abruptly by oral opening. The glottis is wide open for voiceless pulmonic stops, and the vocal folds at the glottis vibrate for voiced pulmonic stops. On the other, glottal stop and glottalized (ejective) sounds are produced by closed glottis.

In the production of glottalized sounds, in addition to some form of constriction (narrowing or closing the air passage) in the vocal-tract system, the excitation air is not pulmonic (not coming from lungs). The larynx with closed glottis moves upward like a piston to generate the excitation air stream. This is the point where the glottalized sounds differ from the pulmonic sounds for the place of articulation. Amharic (the official language of Ethiopia) has five ejective sounds: four stops, /p'/, /t'/, /ʧ'/, /k'/ and one dental fricative /s'/. Each of these five sounds has corresponding voiced and voiceless pulmonic conjugates (Leslau 1995), i.e., sounds with the same place of articulation. Table 6 gives the list of ejective sounds in Amharic along with their corresponding voiced and voiceless pulmonic conjugates according to the place of articulation. In the case of ejective fricative (/s'/), there is an oral constriction (producing frication noise) instead of the oral closure as in the case of ejective stops. Since the acoustic characteristics of ejective sounds differ from the corresponding voiced and voiceless pulmonic sound conjugates, mainly in the source of excitation, epoch-based analysis is useful, in addition to the spectral or spectrographic analysis.

The most frequently used features for analysis of ejective sounds are: VOT, closure duration, overall duration, BSE (burst spectral entropy), ART (amplitude rising time) of the post-target vowel and mean cross-correlation coefficients (MXCC). To measure the VOT and closure duration of the ejective sounds and other voiceless pulmonic stops, it is necessary to determine the instant of burst release and the onset of glottal activity. It is difficult to obtain these instants precisely from the waveform or from the spectrogram (wide-band or narrow-band) due to burst and frication noise, and also due to background noise. The precise location of the onset of the glottal activity is obtained using the zero-frequency filter output, the instant of burst release can be located using the normalized residual computed from the LP analysis, which is relatively low in the silence region than in the burst region. The place of articulation, however, is obtained by using the spectral entropy measured at the output of the burst.

The spectral entropy and MXCC are computed as follows:

Calculate the spectral energy $E[k]$ of a frame of speech using

$$E[k] = \left| \sum_{n=0}^{N-1} x[n] e^{\frac{-j2\pi}{N} kn} \right|^2, \tag{6}$$

where N is the length of the analysis frame, and $k = 0, 1, 2, 3, ..., K-1$ are the sample points in the frequency domain.

The spectral entropy P of a frame of speech is given by

$$P = -\sum_{k=0}^{K-1} \frac{E[k]}{\sum_{l=0}^{K-1} E[l]} \log \left(\frac{E[k]}{\sum_{l=0}^{K-1} E[l]} \right), \tag{7}$$

where K is the number of samples in the frequency domain.

Table 6. Ejective sounds along with their corresponding voiced and voiceless pulmonic conjugates in Amharic. (* Stop sounds found only in loan words.)

Manner	Stops				Fricative
	Labial	Dental	Alveolar	Velar	Dental
Voiced	/b/	/d/	/ʤ/	/g/	/z/
Voiceless	/p/*	/t/	/ʧ/	/k/	/s/
Ejective	/p'/*	/t'/	/ʧ'/	/k'/	/s'/

The mean cross-correlation coefficient γ of the LP residual signal $e[n]$ between two successive cycles for the first three pitch periods of a vowel.

$$\gamma = \frac{1}{3}\sum_{k=1}^{3} \frac{\sum_{m=0}^{M-1} e[m]e[m+N_k+l]}{\sqrt{\sum_{m=0}^{M-1} e^2[m] \sum_{m=0}^{M-1} e^2[m+N_k+l]}}, \tag{8}$$

where $\mathrm{M} = \min(N_k, N_{k+1})$, and N_k and N_{k+1} are the number of samples in successive cycles, l is the small relative shift in number of samples that may be necessary to get the maximum cross-correlation coefficient, and $k = 1, 2, 3$.

Figures 11 and 12 show the spectrogram, waveform, ZFF signal and the normalized residual of the LP analysis for a word with voiceless pulmonic stop and ejective, respectively. From the figures it is obvious that the burst release can be detected from the normalized residual plot by its prominence in amplitude, and the onset of glottal activity from the ZFF signal by the appearance of periodicity. The time interval between the burst release and the onset of glottal activity is the VOT (marked as *V* in figures 11c and 12c). The closure duration (marked as *C*) is measured from the last significant excitation of the preceding vowel to the burst release.

These parameters can also be derived from the waveform and spectrogram, as marked in figures 11a and 12a. We call these parameters as those derived from *spectral* methods, whereas the parameters derived from the normalized residual and the ZFF signal are referred to as derived from *nonspectral* methods. It was shown that the nonspectral methods are more accurate than the spectral methods (Worku 2010).

It is found that ejective stops have less duration of frication than their voiceless pulmonic conjugates. This may be because the amount of air for excitation is limited to the volume of the oral cavity, unlike in the voiceless consonants, where the air comes from the lungs. As can be seen in figures 11 and 12, the amplitude rising time of the vowel following the ejective sounds is longer than for the vowel following the voiceless pulmonic stops.

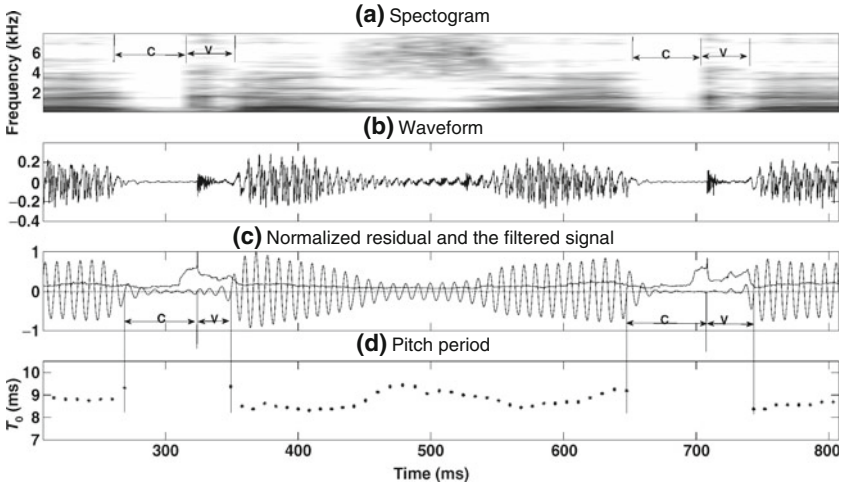

Figure 11. Illustration for voiceless pulmonic stops. (**a**) Wide-band spectrogram, (**b**) waveform, (**c**) filtered (ZFF) signal and normalized residual of the LP analysis, and (**d**) pitch period contour for the part (boldface) of the word **/zəkəzəkə/** (Worku 2010).

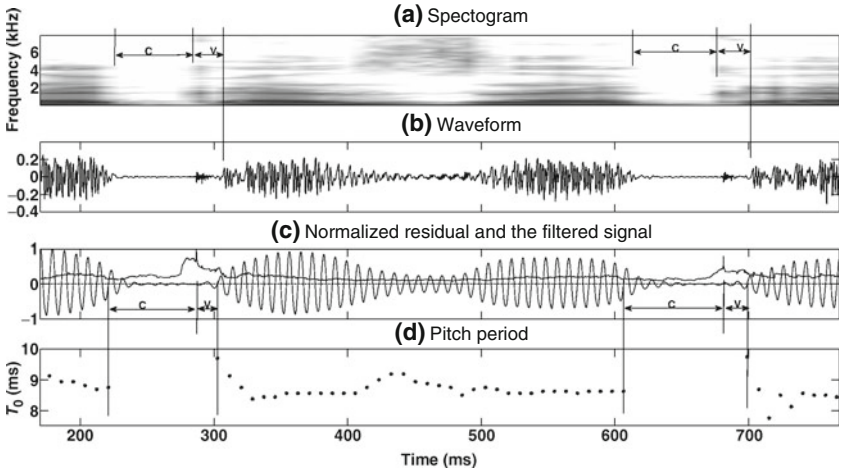

Figure 12. Illustration for ejectives. (**a**) Wide-band spectrogram, (**b**) waveform, (**c**) filtered (ZFF) signal and normalized residual of the LP analysis, and (**d**) pitch period contour for the part (boldface) of the word **/zək'əzək'ə/** (Worku 2010).

All ejective sounds are found to be followed by a creaky (irregular periodicity) vowel, which is the result of dynamics of the vocal folds to change its state from complete tight closure to a relaxed vibration or normal phonation (Stevens & Hajek 2004; Vassiére 1997). Creaky phonation is characterized by having high or irregular pitch period (Dilley *et al* 1996; Gordon & Ladefoged 2001). As can be seen from figures 11c and 12c, the pitch contour of the vowel following /k'/ begins with more irregular pattern than the vowel following /k/. Irregularity of a signal can also be measured by the cross-correlation coefficients of the waveforms in successive pitch periods.

The patterns of glottal stop are different from the supra-laryngeal consonants, in the sense that the features of the vowel preceding the glottal stop spread across it into the following vowel, whereas the spreading does not occur for oral stop consonant (Borroff 2007). If the two flanking vowels are different, then the formant transitions are caused by the place of the oral consonant (C), whereas the transition is caused by the following vowel in the $V_1?V_2$ for the glottal stop. Therefore, the glottal stop information cannot be inferred from the formant contours.

The voice source is aperiodic, and somewhat irregular, in successive periods during the glottal stop region, compared with voice source in the vowel region. The aperiodicity is measured from the successive positive zero crossing intervals in the ZFF output of the LP residual. The normalized cross-correlation coefficient of successive segments is used as a measure of dissimilarity between successive segments.

The dissimilarity is less in the vowel regions compared to the dissimilarity in the glottal stop region (figure 13d). The successive glottal pulse intervals in the vowel regions is more steady compared to the intervals plot in glottal stop region (figure 13e). It should be noted that the formant contours are steady in the glottal stop region also (figure 13f), which shows that the vocal-tract system parameters do not show the presence of the glottal stop.

Creaky voice is a result of laryngealization (where the vocal folds are held stiff and vibration is partially inhibited). Laryngealization can cause the arytenoid cartilages and ligament portion alternating with high and low amplitudes, which is perceived as creaky sound. The irregularity in the pitch period contour and the low values of the normalized cross-correlation coefficient clearly bring out the nature of the creaky voice as shown in figure 14.

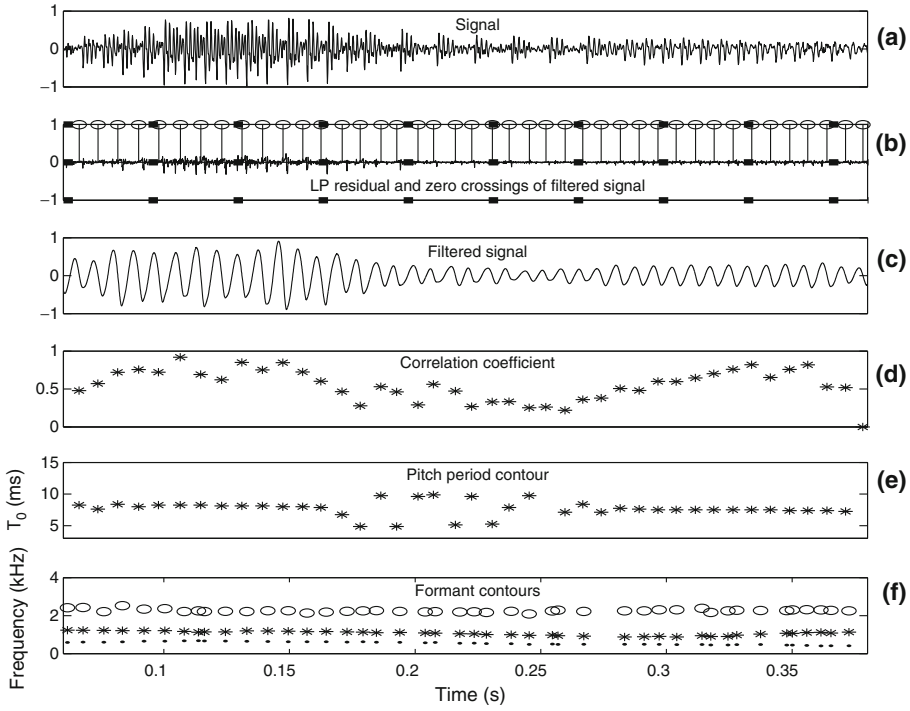

Figure 13. Analysis of glottal stop in V?V context: (**a**) speech signal, (**b**) LP residual and zero crossings of filtered (ZFF) signal, (**c**) filtered (ZFF) signal derived from LP residual, (**d**) normalized correlation coefficient, (**e**) contour of pitch period (T_0) and (**f**) formant contours (Worku 2010).

3.6 *Characterization of loudness of speech*

Several measures of loudness of speech have been proposed in the literature, based on physiological characteristics of speech production and acoustic characteristics of speech signal (Monsen & Engebretson 1977; Gauffin & Sundberg 1989). Acoustic correlates of loudness include the features of the glottal wave (such as closed quotient, open quotient and closing quotient), and spectral features derived from the speech signal (such as spectral tilt, measure of spectral energy in the high-frequency region relative to that in the low-frequency region, harmonic richness, and sharpness of spectral peaks at formant frequencies). It is observed from the EGG signals of soft and loud speech utterances, that the abruptness of closing phase in the glottal cycle is higher in the loud speech, compared to that in the soft speech. This indicates that the impulse-like nature of the glottal excitation plays an important role in the perception of loudness of speech. The objective is to derive a feature in the acoustic speech signal, which represents the impulse-like nature of glottal excitation.

Figure 15 shows segments of voiced speech, chosen from soft, normal and loud utterances of a speaker. The impulse-like excitation, as observed from the LP residuals of the speech segments, is more spread out in time for the soft utterances (figure 15d), compared to the normal (figure 15e) and the loud (figure 15f) utterances. The impulse-like nature of the glottal excitation can be observed clearly from the HE of the LP residual, shown in figures 15g, h, and i for soft, normal, and loud utterances, respectively.

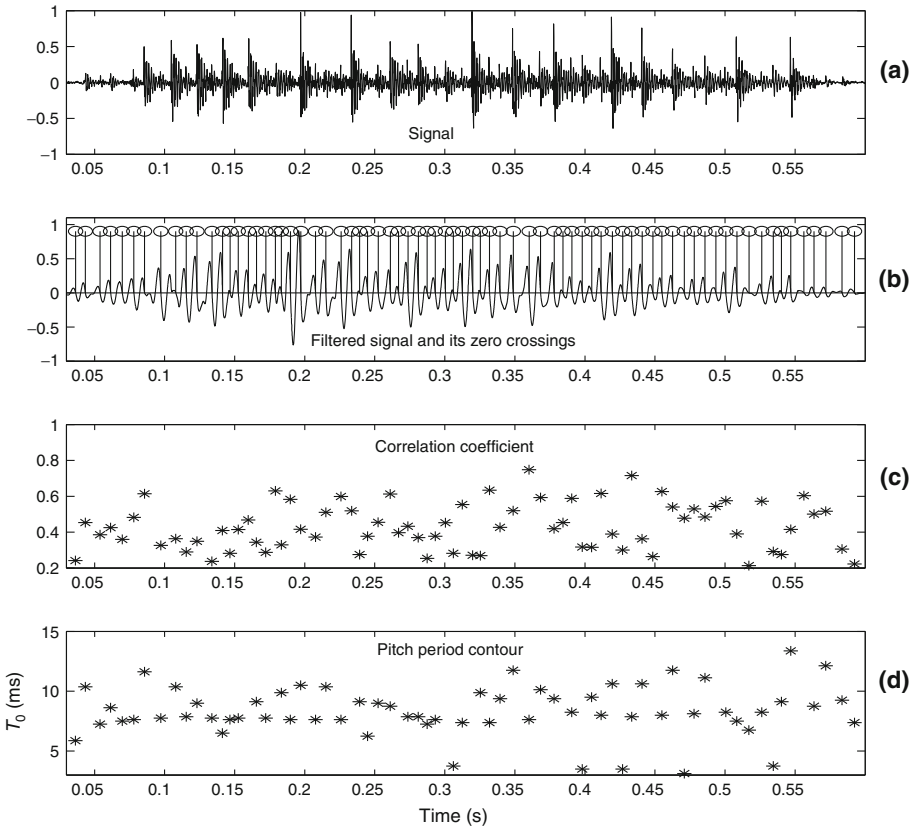

Figure 14. Analysis of creaky utterance 'had'. (**a**) speech signal, (**b**) filtered (ZFF) signal derived from LP residual, and its zero crossing points, (**c**) cross-correlation coefficients of successive pitch period of the LP residual and (**d**) pitch period (T_0) contour (Worku 2010).

A measure of the strength of excitation can be derived from a short segment (about 3 ms) around the instant of impulse-like excitation. To represent the sharpness in the HE of the LP residual, a feature called strength of excitation is defined as $\eta = \frac{\sigma}{\mu}$, where μ denotes the mean of the samples of the HE of the LP residual in a segment around the instant of impulse-like excitation, and σ denotes the standard deviation of the samples. The ratio between standard deviation and mean is also defined in statistics, where it is known as the coefficient of variation. Figure 16 shows an example of the distribution of η for two female and two male speakers, for soft and loud utterances. From figures 16a, c and d, it is observed that the distribution of η does discriminate between the soft and loud utterances of the speakers. The degree of discrimination is not significant in figure 16b. The degree of discrimination, or separation between the distributions of η for soft and loud utterances, is speaker dependent.

3.7 *Analysis of Lombard effect speech*

The speech produced by a person depends on several factors, which include the environment and the auditory self-feedback of the speech of his/her own voice. Adverse environment not only

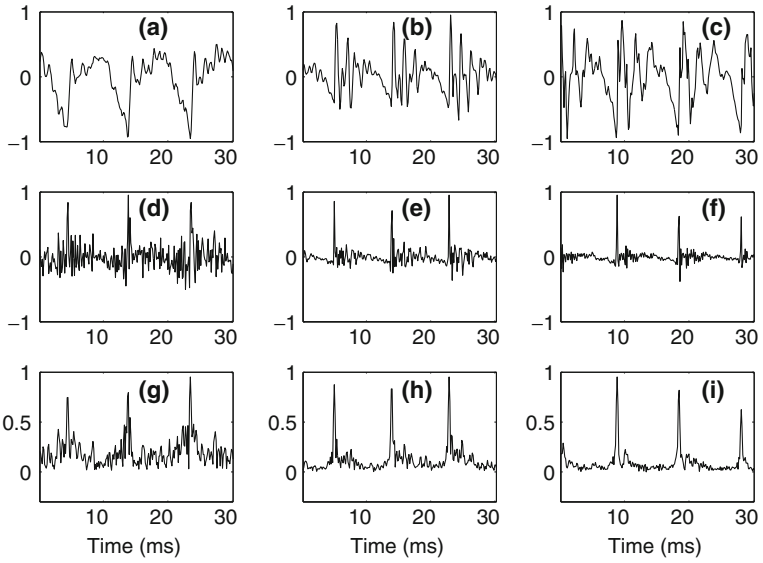

Figure 15. Illustration of the nature of excitation in soft, normal and loud utterances. Speech segments in (**a**), (**b**), and (**c**) belong to soft, normal, and loud utterances, respectively. The segments correspond to the vowel '/a/' in the word 'party' in the sentence, 'She has left for a great party today'. Figures (**d**), (**e**), and (**f**) show the LP residual for the signals in (**a**), (**b**), and (**c**), respectively, while the figures (**g**), (**h**), and (**i**) show the HE of the corresponding LP residuals (Guruprasad & Yegnanarayana 2009).

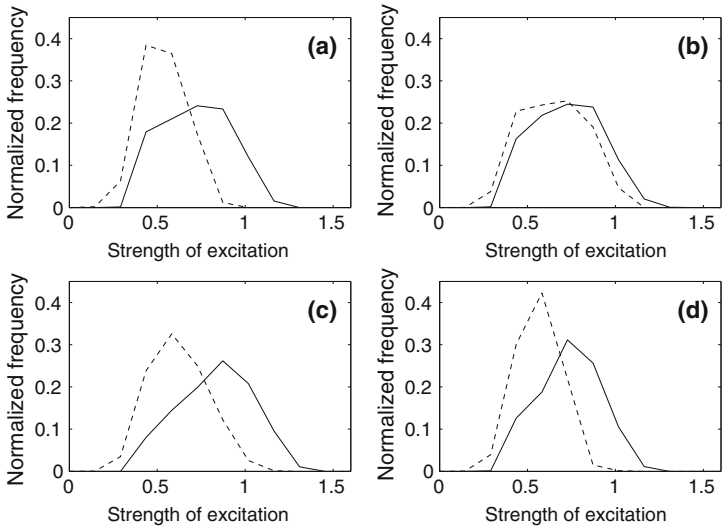

Figure 16. Distribution of the strength (η) of excitation for four speakers. In each case, the broken and the solid lines correspond to soft and loud utterances, respectively. Figures (**a**) and (**b**) correspond to two female speakers, while (**c**) and (**d**) correspond to two male speakers (Guruprasad & Yegnanarayana 2009).

corrupts the speech signals by additive noise, but they also affect the self-feedback of the speech to the person. Lack of self-feedback also affects the articulatory movement in the speech production process, resulting in speech which the listener perceives as not normal. The speaker tries to adjust the articulatory and acoustic parameters to produce speech as intelligence as possible to the listeners. This psychological effect on speaker producing speech in the presence of noise is termed as Lombard effect (Lombard 1911). The Lombard effect not only affects the intelligibility in speech communication, but it also affects the performance of automatic speech and speaker recognition systems.

It is interesting to analyse the Lombard effect speech in terms of features of excitation source in speech production, as the time varying excitation changes significantly under the influence of external feedback. The excitation features considered are the fundamental frequency F_0 (pitch), the strength of the impulse (ϵ) at the epoch and a measure of loudness. An increase in F_0 is observed for the Lombard effect speech compared to the normal speech. The strength of impulse decreases for Lombard effect speech compared to normal speech.

The effect of increased loudness due to Lombard effect cannot be observed in the distribution of the values of the loudness measure (η) proposed in section 3.6. The perceived loudness of the Lombard effect speech not only depends on the sharpness of the peaks in the HE around the epochs, but also on the fundamental frequency (F_0). A measure of perceived loudness (β) is proposed to reflect this (Bapineedu 2010). The new measure β is given by.

$$\beta = \eta \times F_0. \tag{9}$$

Figure 17 shows the β contours of normal speech and for Lombard effect speech, which show the increased loudness for Lombard effect speech in some regions.

3.8 *Analysis of laugh signals*

Laughter is a nonverbal vocalization that occurs often in continuous speech. The vocalized expression of laughter varies across gender, individuals and context. Despite its variability, laughter is perceived naturally by humans. Although laughter has very distinct perceivable

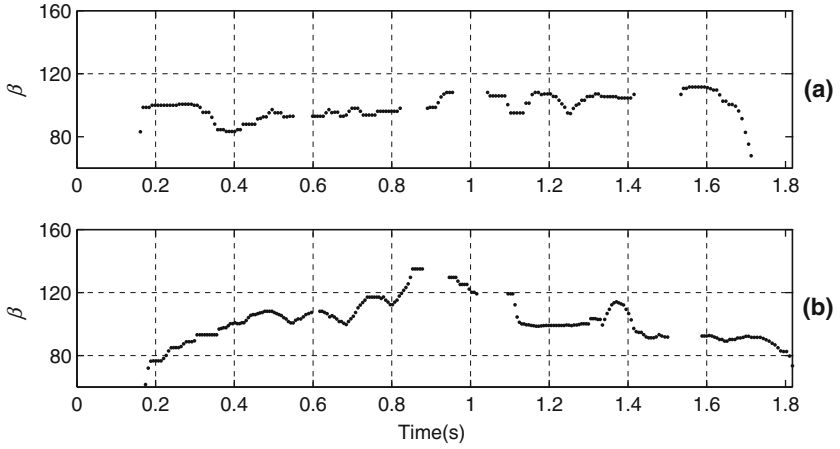

Figure 17. β contours for (**a**) normal speech, and (**b**) Lombard effect speech, for an utterance of the sentence, 'Regular attendance is seldom required' (Bapineedu 2010).

pattern, its production is not guided by any rules of grammar as in the case of speech. It is typically produced by series of sudden bursts (outflows) of air, keeping the vocal tract in a neutral position. The main difference between speech and laughter is that normal speech does not disrupt breath, whereas laughter may (Wallace 2007). In normal speech, major role is played by the articulators of the vocal-tract system, whereas in the case of laughter, major role is played by the lungs and the vocal folds.

Laughter can be broadly divided into two types based on its glottal activity: (i) voiced laughter and (ii) unvoiced laughter. In a voiced laughter, the air bursts flow mostly through the mouth, and in some cases it may even pass through the nose. There will be more air pressure build up in lungs, as a result of which the vocal folds may vibrate in a different way. Further, since there is more air flow, there will be turbulence generated within the vocal folds, and hence the signal may be breathy when compared to normal speech. In unvoiced laughter there is no voicing, the durations of the calls will be less, and there will be more damping when compared to voiced laughter.

Since the vocal tract is in the neutral position while laughing, we hypothesize that most of the laughter characteristics occur due to variations in the excitation. The laughter signals are analysed at two levels: (i) call level and (ii) bout level. Based on this kind of analysis, the features can be categorized into two groups: call level features and bout level features. The call level features are used to capture the call level patterns (variations within a call) of laughter, whereas bout level features are used for capturing the high-level repetitive structure (patterns between calls) of laughter.

The source and system characteristics of laugh signals at call level are analysed using features such as pitch period (T_0), strength of impulse (ϵ) at epochs, amount of breathiness, call durations and some parameters derived from them. It is observed that pitch frequency for laughter is more than that for normal speech (Bachorowski *et al* 2001). The general pattern that is observed in the pitch period within a call is that it starts with some value, decreases slightly to some minimum, and then increases rapidly to a high value as shown in figure 18b. The main issue here is extracting the pitch period accurately.

Since there is large amount of air pressure build up in the case of laughter, (as large amounts of air is exhaled), the closing phase of the vocal folds is very fast. This will result in an increase in the strength of impulse (ϵ) at epoch. Figure 18c shows this general trend of ϵ in the calls within a bout. The pitch period contour of laughter has a unique pattern of rising rapidly at the end of a call. So, the slope of the pitch period contour is used to capture this pattern.

As in the case of the pitch period, the strength of impulse at epochs also changes rapidly. It rises rapidly to some maximum value and again falls at the same rate. Hence, the slope of the normalized strengths are used to capture this pattern. Using the features in the excitation at call and bout levels, it is possible to spot laughter segments in continuous speech as shown in Kumar (2010).

3.9 *Analysis of speech produced at different speaking rates*

When humans modify speaking rate, they do not perform a simple expansion or compression of the speech signal. To maintain the intelligibility and naturalness of the speech, they modify some of the characteristics of the speech production mechanism in a complex way. This causes the acoustic features extracted from the speech signal to change in a complex way. These changes affect the performance of speech systems such as speech recognition and speaker recognition. Most of the studies on the effect of speaking rate on acoustic features focus on features at segmental and suprasegmental levels. Epoch-based analysis is useful to study the effect of

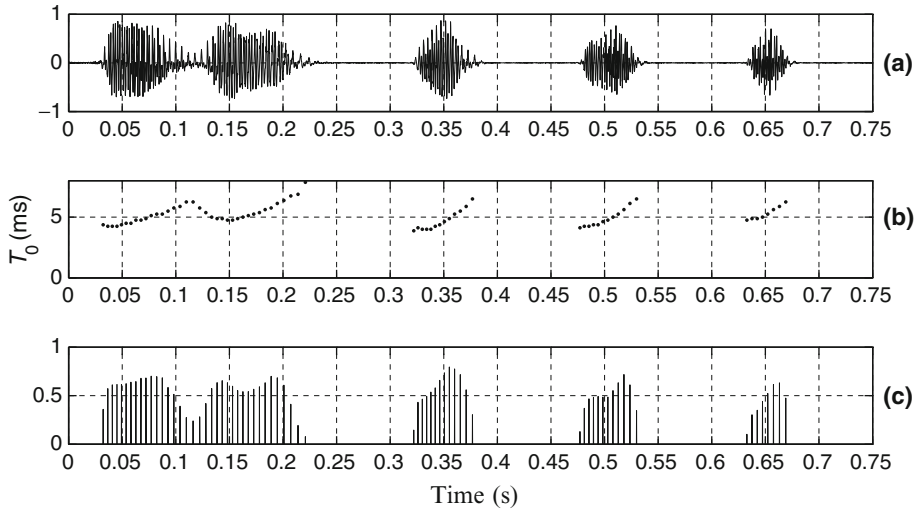

Figure 18. (**a**) Waveform of a voiced laughter. (**b**) Pitch period contour derived from epoch locations. (**c**) Strength of impulses at epochs derived from the zero-frequency filtered signal (Kumar 2010).

speaking rate on features at the subsegmental level. The features are instantaneous F_0, strength of impulse at epoch, perceived loudness, and their distributions at different speaking rates. For this study, the speech material consists of 10 English sentences uttered by 25 male speakers at three different speaking rates, namely, fast, normal and slow.

Figure 19 shows the distributions of instantaneous F_0 for four male speakers, chosen at random from the set of 25 speakers. The distributions of the instantaneous F_0 for four speakers are examined to illustrate differences among individuals, indicating the speaker-specific nature of these variations. It is observed that the distribution of instantaneous F_0 does discriminate between fast, normal, and slow utterances of the speakers, although the amount of discrimination is speaker-dependent. For speakers in figures 19a and b, there is good discrimination between distributions of the instantaneous F_0 for fast, normal and slow utterances. Discrimination can be observed from the mean of the distributions, and from their spreads. For the speaker in figure 19c there is very little difference between distributions of fast and normal utterances, but some discrimination can be seen between the distributions of slow and normal utterances. The distributions shown in figure 19d are very close to each other. Some speaker-specific characteristics can be inferred from the distributions shown in figure 19. For a speaker with a naturally fast speaking rate, the distinction between instantaneous F_0 of his/her fast and normal speech will be less. Similarly, for a speaker with a naturally slow speaking rate, the instantaneous F_0 of slow and normal speech are similar. Some speakers are able to produce speech at three different speaking rates while maintaining intelligibility and naturalness. In most cases, speech uttered in at least one of the non-normal (fast or slow) speaking rates showed significant difference from the speech uttered in normal and the other nonnormal (slow or fast) speaking rate.

Figure 20 shows the distributions of ϵ for the four male speakers (same as used in figure 19). From figure 20, it is observed that ϵ does vary with speaking rate. The degree of variation is speaker-dependent. The general trend observed across all the speakers is that $\bar{\epsilon}$ (where $\bar{\epsilon}$ denotes the mean value of ϵ) of fast speech is less than that of normal and slow speech. For some speakers, $\bar{\epsilon}$ of normal speech is less than that of slow speech (figures 20b and d), whereas for some

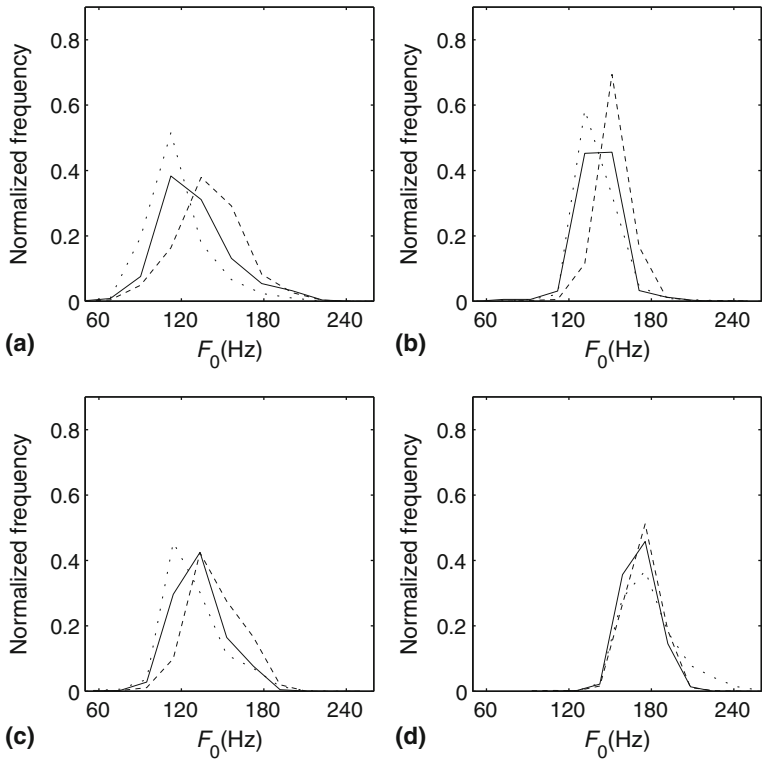

Figure 19. Distributions of instantaneous F_0 for four male speakers. In each case, the solid ('—'), the dashed ('- - -'), and the dotted ('· · ·') lines correspond to normal, fast and slow utterances, respectively (Reddy 2010).

others $\bar{\epsilon}$ of slow speech is less than that of normal speech (figures 20a and c). It should be noted that this is a speaker-specific property. The spread of distribution of the values of ϵ for fast speech is less than that of normal and slow speech. This implies that the variation of ϵ is less for fast speech compared to slow and normal speech. For speakers in figures 20a, b and c, there is good discrimination between the distributions of ϵ of fast, normal, and slow speech. For the speaker in figure 20d, the discrimination is very less.

It is observed that the loudness measure η varies very little with speaking rate, even though the perceptual loudness scores show that fast speech is perceived to be louder than normal speech (Reddy 2010).

3.10 *Processing of speech collected at a distance*

Speech signal collected at a distance (>50 cm) from a speaker is significantly different from that collected close (5–10 cm) to the mouth of the speaker, due to reduction in the amplitude of the direct component with distance, and due to effects of reverberation and noise. Hence, extraction of fundamental frequency of voicing from distant speech signals is a challenging task. The robustness of the impulse-like excitations in voiced speech can be exploited to extract the fundamental frequency from distant speech signals. The key idea is that the impulse-like excitations in the direct component are relatively stronger than those in the reverberant components.

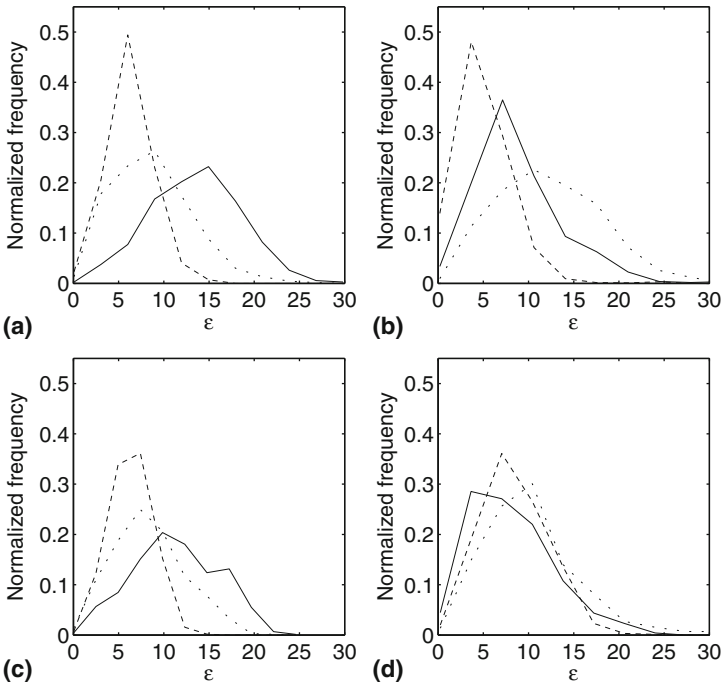

Figure 20. Distributions of strength of excitation (ϵ) for four male speakers. In each case, the solid ('—'), the dashed ('- - -'), and the dotted ('· · · ') lines correspond to normal, fast and slow utterances, respectively (Reddy 2010).

The speech signal is filtered through a cascade of resonators located at zero-frequency (Murty & Yegnanarayana 2008). The ZFF signal preserves the information specific to the fundamental frequency of speech. Figure 21a shows a voiced segment of distant speech signal. The corresponding ZFF signal is shown in figure 21b. The information of the pitch periodicity is seen more clearly in the ZFF signal (figure 21b) than in the distant speech signal (figure 21a), despite the presence of spurious zero crossings in the ZFF signal. The ZFF signal is free from the effects of resonances of the vocal tract. Even though the ZFF signal (figure 21b) consists of some spurious zero crossings, it is observed that the fundamental frequency is the strongest component in the short-time spectrum of the ZFF signal. The location $\tilde{F}_0$ of peak the in the magnitude of the short-time spectrum of the ZFF signal provides an estimate of the fundamental frequency. An all-pole filter is constructed, such that the location of the pole is chosen corresponding to $\tilde{F}_0$. The ZFF signal is passed through the all-pole filter, resulting in a signal which is nearly free of the influence of spurious zero crossings in voiced regions (figure 21c). The reciprocal of the time interval between successive positive-to-negative zero crossings in the refined ZFF signal is hypothesized as the instantaneous fundamental frequency. Thus, this process helps in deriving the instantaneous fundamental frequency accurately (Guruprasad & Yegnanarayana 2010).

Performance of the ZFF method is evaluated on SPEECON database, which consists of speech signals collected in three different environments, namely, car interior, office and living rooms (Iskra *et al* 2002). The speech signals are collected using a head-mounted close-talk microphone, and simultaneously, using microphones placed at distances of 30 cm, 1 m and 2–3 m from the speaker. These four cases are denoted by C_0, C_1, C_2 and C_3, respectively. Gross error is used to

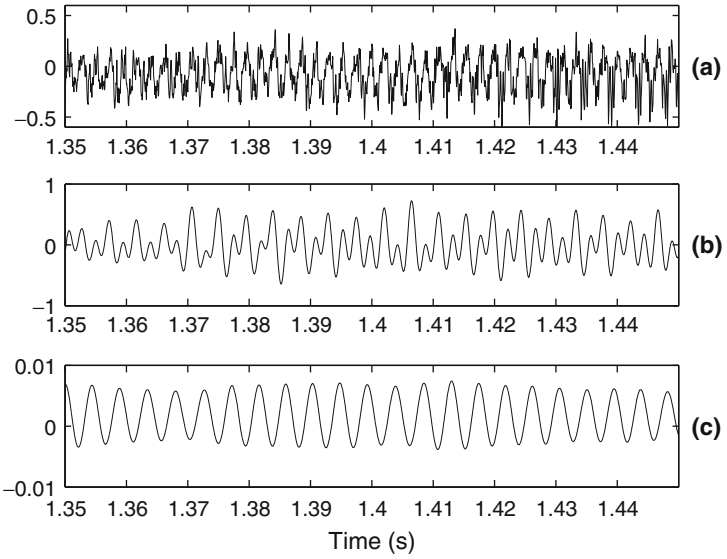

Figure 21. (**a**) Speech signal collected at a distance. (**b**) The filtered (ZFF) signal. (**c**) The refined filtered (ZFF) signal (Guruprasad 2010).

evaluate the accuracy of extraction of the fundamental frequency. It is defined as the percentage of voiced frames for which the extracted value of the fundamental frequency deviates from the reference value by more than 20%. In addition, the mean and standard deviation of the absolute value of the difference between the extracted and the reference values of fundamental frequency are also used for evaluation. Performance of the ZFF method is compared against those of some state-of-the-art methods, which are based on time-domain and frequency-domain processing of speech. These include autocorrelation method (Boersma 1993), crosscorrelation method (Goldberg & Riek 2000), robust algorithm for pitch tracking (Talkin 1995), subharmonic summation method (Hermes 1988), subharmonic-to-harmonic ratio method (Sun 2002), and fundamental frequency estimator (de Cheveigne & Kawahara 2002). Table 7 lists the performance of the ZFF method and those of the existing methods, for the task of extraction of the fundamental frequency. The performance is listed for speech signals collected in office environment. The ZFF method is denoted by ZFFM, which is based only on the interval between successive impulse-like excitations. Table 7 shows that the ZFFM performs better than the existing algorithms. The mean error of ZFFM is greater than that of the other methods, due to inclusion of more frames in the computation of the mean error. The method is also robust for speech signals collected from car interior and living room environments (Guruprasad & Yegnanarayana 2010).

3.11 *Pitch extraction from multispeaker data*

The signal collected by a microphone in a multispeaker environment is a mixture of speech signals from several speakers. Pitch extraction from multispeaker speech signals is a challenging task, as the pitch periods from all the speakers overlap, making it difficult to observe the individual pitch periods. Figure 22a and b show the speech signals collected by a pair of microphones when two persons (one male and one female) are speaking simultaneously. It is difficult to observe the pitch periods corresponding to the speakers from any of the two signals. Even

Table 7. Performance of the proposed method for estimation of fundamental frequency in close-speaking and distant speech signals, collected in office environment. Performance of six existing methods is also listed for comparison. For each distance, the least values of gross error, mean error and standard deviation among the different methods are indicated in boldface (Guruprasad & Yegnanarayana 2010). The abbreviations in the table corresponds to methods: AC – autocorrelation, CC – crosscorrelation, SHS – subharmonic summation, SHRP – subharmonic-to-harmonic ratio, RAPT – robust algorithm for pitch tracking, YIN – fundamental frequency estimator, and ZFFM – zero-frequency filtering method.

	GE (%)				M (Hz)				SD (Hz)			
	C_0	C_1	C_2	C_3	C_0	C_1	C_2	C_3	C_0	C_1	C_2	C_3
AC	6.67	8.14	10.28	32.74	3.12	3.32	3.37	6.09	4.77	5.10	5.29	7.49
CC	6.68	8.54	10.83	33.75	**2.51**	**2.78**	3.24	7.00	**4.15**	4.61	5.09	7.99
SHS	14.06	16.56	18.91	45.63	2.94	3.00	**2.98**	**5.14**	4.20	**4.40**	**4.53**	**6.81**
SHRP	8.98	10.96	20.15	61.17	3.32	3.38	3.37	5.43	4.54	4.70	4.74	6.86
RAPT	9.02	10.85	16.26	44.81	2.91	3.13	3.54	6.98	4.65	4.82	5.26	7.92
YIN	7.03	10.25	15.65	38.55	4.50	4.46	4.31	5.56	5.65	5.72	5.76	6.91
ZFFM	**4.68**	**6.94**	**9.54**	**18.98**	4.99	5.20	5.43	9.53	4.60	5.27	5.48	10.25

the autocorrelation of a frame (30 ms) of the two-speaker signal is not likely to yield two unambiguous peaks corresponding to the pitch periods of both the speakers.

The characteristics of excitation around the epochs, and the robustness of the relative spacing of the epochs in the speech signals collected at a pair of microphones can be exploited for identifying the epoch locations corresponding to a given speaker. Let $g_1[n]$ and $g_2[n]$ be the preprocessed HE ($h_1(n)$ and $h_2(n)$) of the LP residuals of the speech signals collected at *Mic-1* and *Mic-2*, respectively, as given in Eq. 10.

$$g_i[n] = \frac{h_i^2[n]}{\frac{1}{2M+1}\sum_{m=n-M}^{n+M} h_i[m]}, \qquad i = 1, 2 \tag{10}$$

By aligning $g_1[n]$ and $g_2[n]$ after compensating for the estimated time-delay ($\hat{\tau}_1$) corresponding to *Spkr-1*, the epochs corresponding to that speaker will be in coherence, whereas the epochs corresponding to *Spkr-2* will be incoherent. By considering the minimum of $g_1[n]$ and $g_2[n-\hat{\tau}_1]$, only the preprocessed HE around the epochs corresponding to *Spkr-1* are retained. It should be noted that this operation of retaining minimum ensures that the preprocessed HE peaks at the epochs of the other speakers are suppressed. The resulting signal is referred as the HE specific to *Spkr-1*. In a similar manner, the signal that retains the HE around the epochs corresponding to *Spkr-2* can be derived. Let

$$h_{sj}[n] = \min(g_1[n], g_2[n - \hat{\tau}_j]), \qquad j = 1, 2, \tag{11}$$

where $h_{s1}[n]$ and $h_{s2}[n]$ are the signals in which the HE around the epochs corresponding to *Spkr-1* and *Spkr-2*, respectively, are retained.

Figure 22 illustrates the extraction of speaker-specific HE from two-speaker speech signals collected using a pair of microphones. Figures 22c and d show the HE of the LP residuals of the two-speaker speech signals shown in figure 22a and b, respectively. The HE consist of the impulse-like excitations due to the epochs of both the speakers. It is difficult to separate the

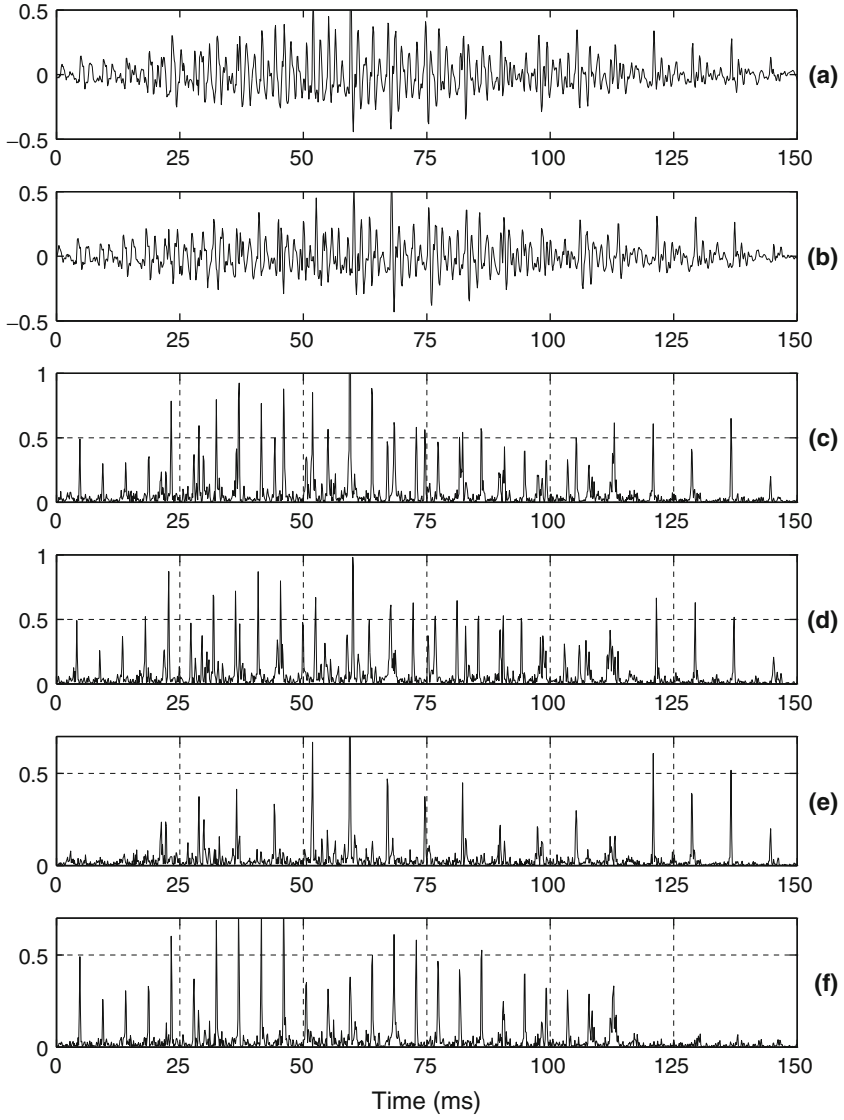

Figure 22. Illustration of extracting speaker-specific HE from two-speaker data collected using a pair of microphones. Speech signal collected at (**a**) *Mic-1* and (**b**) *Mic-2*. HE of LP residual of (**c**) *Mic-1* signal, and (**d**) *Mic-2* signal. Speaker-specific HE of (**e**) *Spkr-1* and (**f**) *Spkr-2*. The time-delays of arrival due to *Spkr-1* and *Spkr-2* are 0.5 ms and −0.625 ms, respectively (Murty 2009).

peaks due to epochs of the individual speakers from any one of them. However, the speaker-specific HE (figure 22e and f), obtained by computing the minimum of the delay-compensated HE, clearly show epochs due to individual speakers.

The speaker-specific HE predominantly contain impulse-like excitations at the epoch locations of the respective speakers. 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 peaks from the speaker-specific HE, that contains large variation among the peak

amplitudes. In order to avoid the difficult task of peak detection, the ZFF approach is employed to detect the impulse-like excitations in the speaker-specific HE. The positive zero-crossings of the filtered HE closely match with the peaks in the speaker-specific HE as shown in figure 23. Figure 23a and b show the speaker-specific HE of *Spkr-1* and its ZFF signal, respectively. Even the low amplitude peaks around 90 ms to 120 ms are correctly detected, while the spurious peaks in between 50 ms to 60 ms are rightly ignored. Similar observations can be made from figure 23c and d, which show speaker-specific HE of *Spkr-2* and its ZFF signal. Hence, the zero-crossings of the signal of the speaker-specific HE can be used to estimate the pitch of the individual speakers.

4. Application of epoch-based analysis

In this section, some practical applications that exploit the impulse-like excitation of the vocal-tract system and the precise locations of the epochs are described. Applications involving time

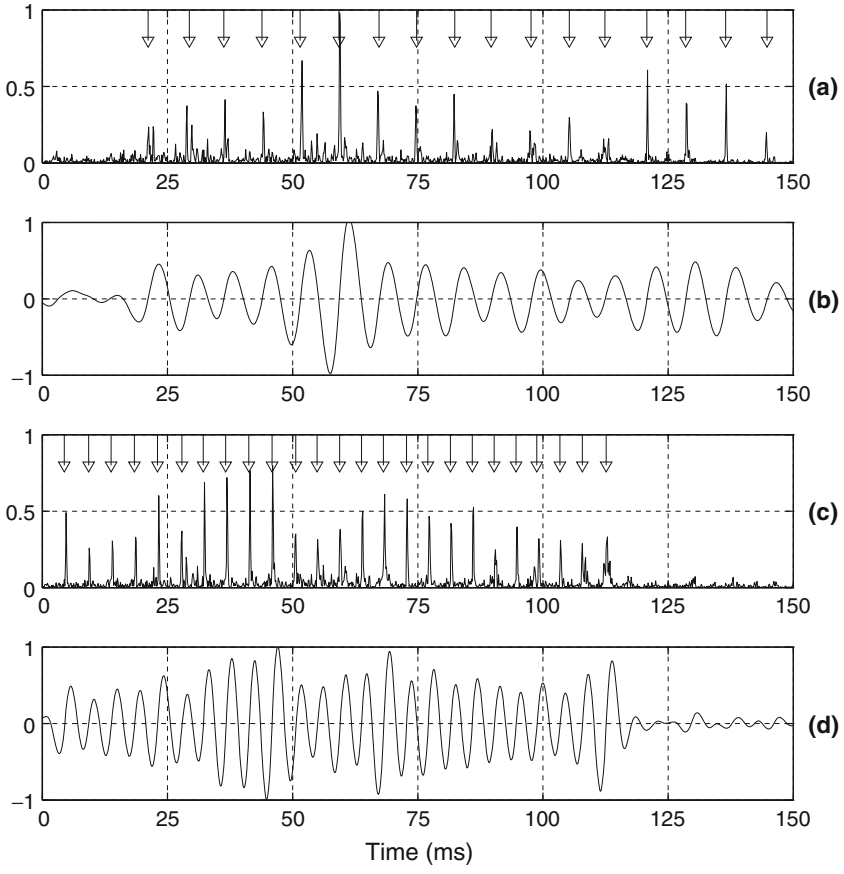

Figure 23. Illustration of epoch extraction from speaker-specific HE using zero-frequency resonator. (**a**) Speaker-specific HE of *Spkr-1* and its (**b**) zero-frequency filtered (ZFF) signal. (**c**) Speaker-specific HE of *Spkr-2* and its (**d**) zero-frequency filtered (ZFF) signal. The arrow marks in (**a**) and (**c**) indicate the epoch locations detected from zero-crossings of ZFF signals in (**b**) and (**d**), respectively (Murty 2009).

delay estimation and speech enhancement are considered in this section (Swamy *et al* 2007; Yegnanarayana *et al* 2005). There are other applications such as pitch synchronous analysis for speaker recognition and enhancement of throat microphone speech, which are not considered here but reported in Joseph *et al* (2009), Murty *et al* (2008) and Reddy *et al* (2010).

4.1 *Time delay estimation and determining number of speakers from mixed signals*

One of the important problems in signal processing is to estimate the number of sources from multisensor data. In multispeaker data, the problem is to determine the number of speakers, and then localize and tract the speakers from the mixed signals received by the sensors. Solutions to these problems are needed, especially for signals collected in a practical environment, such as a room with background noise and reverberation.

In a multispeaker multimicrophone scenario, assuming that the speakers are stationary with respect to the microphones, there exists a fixed time-delay of arrival of speech signals (between every pair of microphones) for a given speaker. The time-delays corresponding to different speakers can be estimated using the cross-correlation function of the multispeaker signals. Positions of dominant peaks in the cross-correlation function of the multispeaker signals give the time-delays due to all the speakers at the pair of microphones. However, in general, the cross-correlation function of the multispeaker signals does not show unambiguous prominent peaks at the time-delays. This is mainly because of the damped sinusoidal components in the speech signal due to resonances of the vocal tract, and also because of the effects of reverberation and noise. These effects can be reduced by exploiting the characteristics of the excitation source of speech. In particular, the speech signal exhibits relatively high SNR and high signal-to-reverberation ratio (SRR) in the vicinity of time instants of significant excitations of the vocal tract. Although some reflected components and noise may also contribute to some high SNR regions, their relative positions will be different in the signals collected at the two microphones. Hence, the coherence of the high SNR regions in the direct components of the signals at the two microphones can be exploited for estimating the time-delay.

In order to highlight the high SNR regions in the speech signal, LP residual is derived from the speech signal using the autocorrelation method of LP analysis (Makhoul 1975). The LP residual removes the second-order correlations among the samples of the signal, and produces large amplitude fluctuations around the instants of significant excitation. The high SNR regions around the GCIs can be highlighted by computing the HE of the LP residual (Ananthapadmanabha & Yegnanarayana 1979).

The cross-correlation function of the HE of the LP residual signals derived from the multispeaker mixed signals is used to determine the number of speakers. Apart from the large amplitudes around the instants of significant excitation, the HE also contains a large number of small positive values, which may result in spurious peaks in the cross-correlation function. The regions around the instants of significant excitation are further emphasized by dividing the square of each sample of the HE by the moving average of the HE computed over a short window around the sample (see Eq. 10).

For multispeaker signals collected using a pair of microphones, $p = 2$. The cross-correlation function $r_{12}[l]$ between the preprocessed HE $g_1[n]$ and $g_2[n]$ is computed as

$$r_{12}[l] = \frac{\sum_{n=z}^{N-|k|-1} g_1[n]g_2[n-l]}{\sqrt{\sum_{n=m}^{N-|k|-1} g_1^2[n] \sum_{n=m}^{N-|k|-1} g_2^2[n]}}, \quad l = 0, \pm 1, \pm 2, \ldots, \pm L, \tag{12}$$

where $m = l$, $k = 0$ for $l \geq 0$, and $m = 0$, $k = l$ for $l < 0$, and N is the length of the segments of the HE. Here, both the vectors are normalized to unit magnitude for every sample shift before computing the cross-correlation. The cross-correlation function is computed over an interval of $2L + 1$ lags, where $2L + 1$ corresponds to an interval greater than the largest expected delay. The largest expected delay can be estimated from the approximate positions of the speakers and microphones in the room. The locations of the peaks with respect to the origin (zero lag) of the cross-correlation function correspond to the time-delays between the microphone signals for all the speakers. The number of prominent peaks should correspond to the number of speakers. However, in practice, this is not always true because of the following reasons: (i) All speakers may not contribute to voiced sounds in the segments used for computing the cross-correlation function. (ii) There could be spurious peaks in the cross-correlation function, which may not correspond to the delay due to a speaker. Hence, we rely only on the delay due to the most prominent peak in the cross-correlation function. This delay is computed from the cross-correlation function of successive frames of 50 ms duration. The delays are computed for each frame shift of 5 ms. Since different regions of speech signal may provide evidence for the delays corresponding to different speakers, the number of frames corresponding to each delay is accumulated over the entire data. This helps in the determination of number of speakers, as well as their respective delays. Thus, by collecting the number of frames corresponding to each delay over the entire data, there will be large evidence for the delays corresponding to the individual speakers.

Speech data was collected simultaneously using two microphones separated by about 1 m in a laboratory environment, with an average (over the frequency range of 0.5–3.5 kHz) reverberation time of about 0.5 seconds. The cross-correlation function of the HE of the LP residuals of the multispeaker signals is used to estimate the time-delays. The percentage of frames for each delay (in ms) for three, four, five and six speakers are shown in figure 24. The locations of the peaks in the histograms correspond to the time-delays due to different speakers. Thus, the number of peaks in the histogram indicates the number of speakers, and the heights of the peaks show the relative prominence of each speaker in the conversation.

4.2 *Speech enhancement in single channel case*

Perceiving information from speech signals collected over severely degraded channels is a difficult task. To increase the comfort level of listening, one can process the speech signal to reduce noise in the nonspeech regions. This is possible if we are able to identify the speech regions. Noise characteristics may be estimated and subtracted from the degraded speech signal. However, in real environments, noise characteristics vary over time. Hence, reliable estimation of noise characteristics is a difficult task. Alternatively, characteristics of speech may be exploited to process the degraded speech. One advantage of using the knowledge of speech is that the characteristics of speech are more predictable compared to that of the noise components (Yegnanarayana & Murthy 2000).

The speech-specific knowledge from the vocal-tract system or the excitation source or both may be used for enhancement. In this study, the knowledge of the excitation source is used for enhancement. The HE of LP residual containing information about the excitation source is derived from the speech signal. One property of the HE of LP residual of the speech signal collected over a severely degraded channel is that the correlation among the samples is high in speech regions and low in nonspeech regions. Autocorrelation analysis may be performed on the HE of the LP residual to estimate the amount of correlation among the samples. For illustration, a 30 ms frame of HE of LP residual computed from a strong-voiced segment of degraded speech,

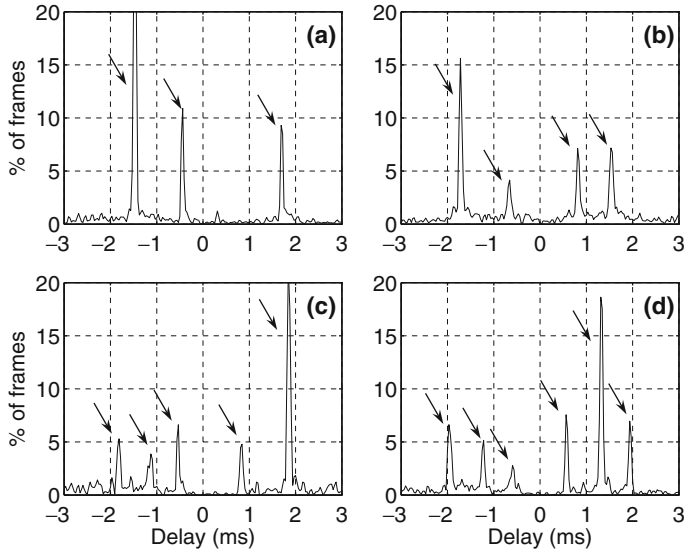

Figure 24. Percentage of frames for each delay in milliseconds for (**a**) three speakers, (**b**) four speakers, (**c**) five speakers and (**d**) six speakers. The arrows indicate the peaks corresponding to different speakers (Swamy *et al* 2007).

and its autocorrelation sequence are shown in figure 25a and b, respectively. The strength of the first peak (after the central peak) in the autocorrelation sequence is an indication of the level of correlation in the frame, which is high in this case. The HE of the LP residual of a 30 ms frame of a segment of a weak-voiced speech segment and its autocorrelation sequence are shown in figure 25c and d, respectively. The strength of the peak is relatively low in this case. Similarly, autocorrelation analysis performed for the HE of the LP residual of a 30 ms frame of nonspeech regions of the signal is also shown in figure 25e and f, respectively. Therefore, the autocorrelation analysis performed on frames of the HE of the LP residual for every sample shift gives an indication of the level of speech at each sample in the degraded signal.

A 10^{th} order LP analysis is performed on the degraded speech signal (see figure 26a) sampled at 8 kHz, to obtain the LP residual. The HE of the LP residual is computed. Autocorrelation is performed on the HE of the LP residual using frames of size 30 ms and frame shift of one sample. For each frame, the strength of the first peak of the autocorrelation sequence, normalized with respect to the central peak, is noted. The normalized peak strength of the autocorrelation sequence, computed for the HE of the LP residual is shown in figure 26c. High values in the normalized peak strength indicate the speech regions. The normalized peak strength sequence is suitably processed using a 500-point Hamming window. A weight function is derived from the smoothed sequence using a nonlinear mapping function in such a way that the samples corresponding to the speech regions are enhanced relative to the samples in the nonspeech regions. The nonlinear mapping function may be given by

$$P_m = \frac{1}{1 + e^{-(P_s - \theta)/\tau}} + \alpha, \tag{13}$$

where P_m is value of the weight function value, P_s is the smoothed peak strength value (normalized in the range 0–1). The parameters $\theta = 0.2$, $\tau = 0.04$ are the slope parameters, and

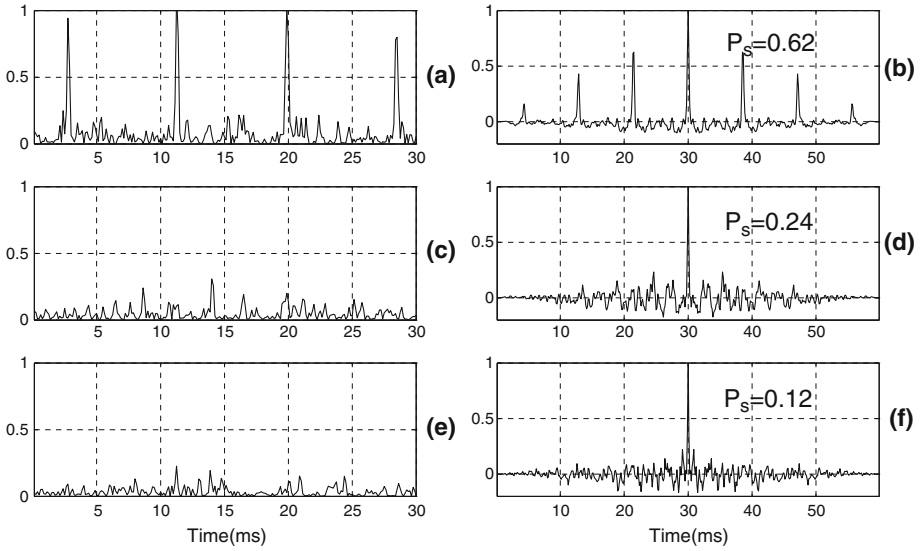

Figure 25. HE of the LP residual of a 30 ms (**a**) strong-voiced frame and its (**b**) autocorrelation sequence, (**c**) weak-voiced frame and its (**d**) autocorrelation sequence, (**e**) nonspeech frame and its (**f**) autocorrelation sequence. P_s indicates normalized strength of the first peak (Prasanna 2004).

$\alpha = 0.05$ is the offset which is the minimum value of the weight function. The weight function derived using the mapping function is shown in figure 26d.

The LP residual of the degraded speech signal is processed using the weight function to produce the modified LP residual (figure 26e). As the samples of the LP residual signal are less correlated compared to the samples of the speech signal, modifying the LP residual may introduce less distortion in the synthesized signal. The enhanced speech signal (figure 26f) is obtained by exciting the time-varying filter using the modified LP residual. The LP coefficients of the filter are derived from the degraded speech signal.

From the narrowband spectrograms of the degraded and the corresponding enhanced speech signals, we can infer that the energy of the frequency components in the speech regions are unaltered, but the signal is attenuated significantly in the nonspeech regions (Prasanna 2004).

4.3 *Multichannel speech enhancement*

When speech is transmitted in an acoustical environment similar to an office room, it will be degraded by background noise and reverberation (Boll 1979; Cheng & O'Shaughnessy 1991; Ephraim & Trees 1995; Flanagan *et al* 1985; Huang & Zhao 1998; Jensen & Hansen 2001; Mittal & Phamdo 2000; Miyoshi & Kaneda 1988; Nemer *et al* 2002; Oh & Viswanathan 1992; Satyanarayana 1999; Scalart & Benmar 1996; Silverman 1987; Subramaniam *et al* 1996; Yegnanarayana *et al* 1997, 1999; Yegnanarayana & Murthy 2000). Multichannel case is more effective for enhancement compared to the single channel case, but requires estimation of time-delays (Flanagan *et al* 1985). A simple method for enhancement in multichannel case is addition of the speech signals, after compensating for the delays. Coherent addition of speech signals from different microphones will provide enhancement mainly against background noise. The improvement in enhancement is directly related to the number of microphones used. For

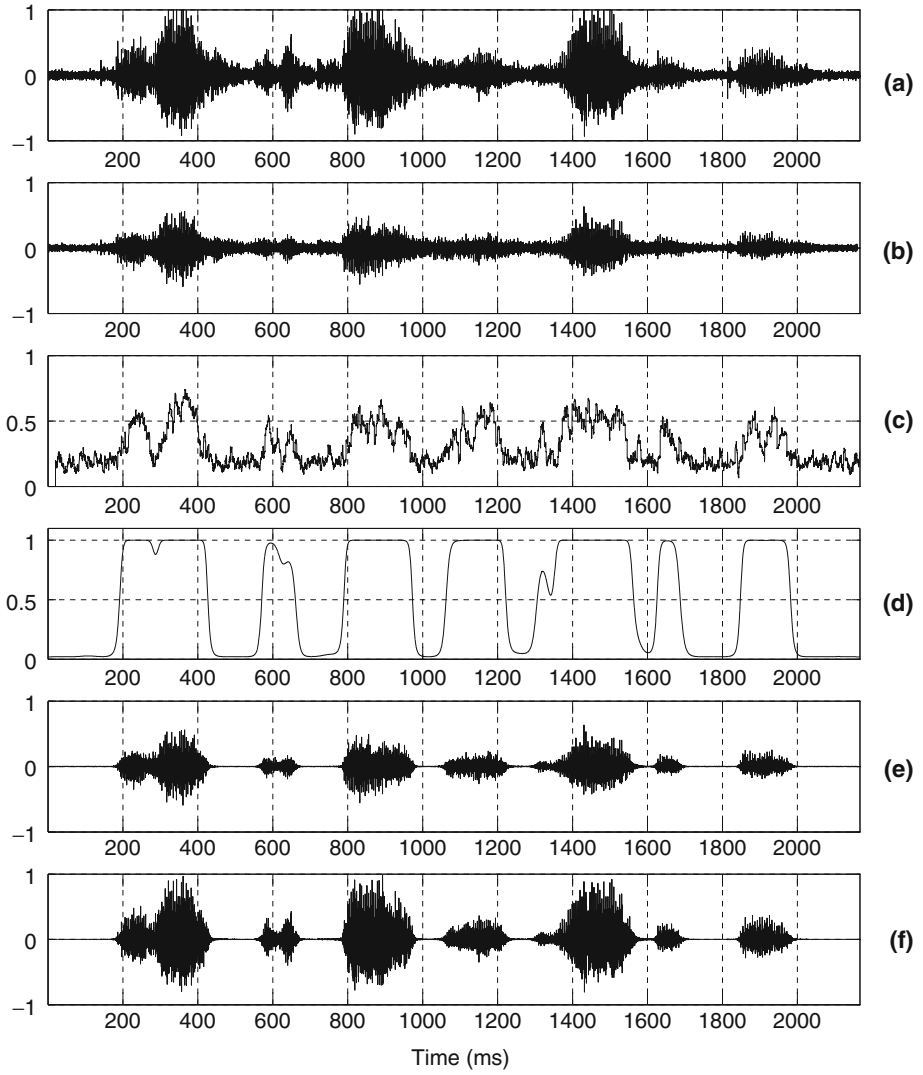

Figure 26. (**a**) Degraded speech signal, and its (**b**) LP residual, (**c**) normalized peak strength values, (**d**) weight function, (**e**) modified LP residual, and (**f**) enhanced speech signal (Prasanna 2004).

achieving significant enhancement, especially due to reverberation, additional processing of the microphone signals is required. In this section, a method for enhancement using the epoch information is described.

Speech was collected from 14 spatially distributed microphones placed in an office room of dimension 3m × 4m × 3m with a reverberation time of about 200 ms. The delay between every pair of microphones is computed using the excitation source information. The coherently-added signal, obtained after compensating for their delays is given by

$$s_{e1}(n) = \frac{1}{N}[s_1(n) + s_2(n - \tau_{12}) + ... + s_N(n - \tau_{1N})], \tag{14}$$

where τ_{1i} is the delay in samples between *mic-1* and *mic-i*.

The coherent addition reinforces speech components and thus reduces the effect of the background noise. However, the reverberant component is still present in the resulting signal. The degree of enhancement achieved at this level depends on the number of microphones used in the coherent addition. For instance, from the enhanced speech signals and their narrowband spectrograms, when signals from 2, 5, 10 and 14 microphones are added, one can observe that there is decrease in the background noise as the number of microphones are increased (Prasanna 2004). It is interesting to note that in the case where signals from 14 microphones are added, even though the effect of background noise is reduced, the reverberation tails are still present in the signal. It is necessary to process the coherently-added speech signal further to achieve enhancement with respect to reverberation.

For each of the microphone signals, the HE of the LP residual is computed. The coherent addition of the HE reinforces the epoch information, whereas the incoherent addition will spread the epoch information. The coherently-added HE exhibits several interesting features. The deviation among the samples of the HE is high in the voiced speech regions. Typically, voiced speech regions in continuous speech have a minimum duration of 50 ms. Hence, by considering a block of 50 ms duration and a shift of one sample, the mean and standard deviation of the coherently-added HE samples in each block are computed. The standard deviation values are normalized with the respective mean values. In the normalized standard deviation plot, the deviation of HE samples is high in the speech regions. Further, the normalized standard deviation is high in the initial portions of the voiced speech regions, and it decreases towards the end of the voiced regions. This is because the initial parts are high SNR regions. Towards the end of the voiced regions, the relative levels of degrading components increase, and hence they correspond to low SNR regions. Another interesting property of the coherently-added HE is that, the samples in each pitch period around the epochs have large deviation compared to the samples away from the epoch. The mean, standard deviation and normalized standard deviation of the coherently-added HE are computed for a block size of 3 ms and a shift of one sample.

A weight function is derived by adding the two (long and short blocks) normalized standard deviation values as shown in figure 27. The weighted residual is used to excite the time varying all-pole filter derived from the coherently-added signal, to synthesize the enhanced speech signal. The clean speech, its degraded version, coherently-added signal from three microphones and the enhanced speech, along with their narrowband spectrograms indicate that the speech signal is enhanced both with respect to background noise as well as with respect to reverberation (Prasanna 2004, p. 70–83).

4.4 *Multispeaker separation*

In a multispeaker environment, the objective is to separate the speech component corresponding to each speaker, while retaining the quality and intelligibility as much as possible. The signal collected by a microphone in a multispeaker environment is a mixture of speech signals from several speakers. Processing speech for enhancement in such conditions is a challenging task, as the speech of the other speakers acts as noise, against which the speech of the desired speaker needs to be enhanced. The difficulty in achieving this enhancement is due to the similarity of the spectral characteristics of the speech signals from different speakers. The difficulty is further compounded by the fact that the spectral characteristics are modified by the response of the room and also by the background noise. The extent of degradation depends on the relative position of the microphone with respect to the speaker, and also on the background noise. The primary

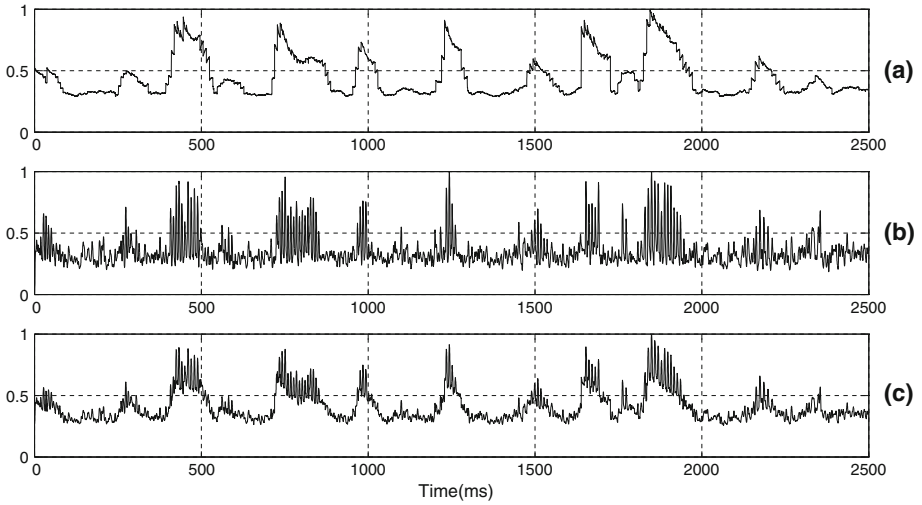

Figure 27. (**a**) Normalized standard deviation plot derived using block size of 50 ms and shift of 1 sample, (**b**) normalized standard deviation plot derived using block size of 3 ms and shift of 1 sample, and (**c**) weight function obtained by adding (**a**) and (**b**) (Prasanna 2004).

causes of degradation are room reverberation, background noise and the distance of the speaker from the microphone.

Most of the multispeaker enhancement methods in the literature involve modification of spectral features representing the vocal tract system (Barros *et al* 2002; Cardoso 1998; Lee & Childers 1988; Mitchell *et al* 1971; Morgan *et al* 1997; Frazier *et al* 1990; Parsons 1976). They use the knowledge of pitch to separate the individual speakers in multispeaker case. Here, a method is described for speaker separation from speech collected over multiple microphones, by exploiting characteristics of the strength of excitation impulses at the epochs, and the robustness of the relative spacing of the GC events in the speech signals collected at different microphones. The HE of the LP residual is used as a representation for the sequence of impulses corresponding to the instants of significant excitation of the vocal tract system. When these sequences are added coherently using the knowledge of the time-delay of each speaker, the strengths of the excitation of the desired speaker are enhanced relative to the strengths of excitation of other speakers. Using the knowledge of the enhanced speaker characteristics in the coherently-added sequence of impulses, a weight function is derived, which in turn is used to derive a modified excitation signal. This modified excitation signal is used to synthesize speech using the vocal tract system characteristics derived from the degraded speech signal. Enhancement in the resulting speech is primarily due to enhancement of the excitation characteristics, which are important perceptually (Prasanna 2004).

The steps in processing the two microphone signals for enhancement are summarized in table 8 (Prasanna 2004).

4.5 *Prosody modification using knowledge of epoch locations*

Prosody modification involves changing pitch and duration of speech without affecting the naturalness. This section gives a method for prosody modification using the instants of significant

Table 8. Steps for speech enhancement in multispeaker environment (Prasanna 2004).

Sl. No.	Description
1	Collect the speech signals (sampling frequency 8 kHz) from two speakers over two spatially separated microphones in a live room.
2	Derive the (10^{th}-order) LP residuals from the speech signals.
3	Compute the HE of the LP residuals.
4	Estimate the time-delays for each speaker using the cross-correlation of the HE.
5	Add the HE using the estimated time-delays to produce the coherently-added HE for each speaker.
6	Derive the weight function using the standard deviation plots from the coherently-added HE.
7	Derive the modified LP residual signal from each microphone signal.
8	Synthesize the enhanced speech for each microphone signal.
9	Coherently add the speech signals of the desired speaker derived from both the microphone signals.

excitation of the vocal tract system during production of speech. The instants of significant excitation correspond to the epochs in the case of voiced speech, and to some random excitation in the case of unvoiced speech.

The method for prosody manipulation makes use of the properties of the excitation source information. The residual signal in the LP analysis is used as an excitation signal (Makhoul 1975). Successive samples in the LP residual are less correlated compared to the samples in the speech signal. The residual signal is manipulated by using resampler either for increasing or decreasing the number of samples required for the desired prosody modification. The residual manipulation is likely to introduce less distortion in the speech signal synthesized using the modified LP residual and LP coefficients (LPCs). The time-varying vocal-tract system characteristics are represented by the LPCs for each analysis frame. Since the LPCs carry the information about the short-time spectral envelope, they are not altered in the described method for prosody modification. LP analysis is carried out over short segments (analysis frames) of speech data to derive the LP coefficients and the LP residual for the speech signal (Makhoul 1975).

There are four main steps involved in the prosody manipulation: (i) Deriving the instants of significant excitation (epochs) from the LP residual signal, (ii) deriving a modified (new) epoch sequence according to the desired prosody (pitch and duration), (iii) deriving a modified LP residual signal from the modified epoch sequence, and (iv) synthesizing speech using the modified LP residual and the LPCs. Figure 28 shows the block diagram indicating various stages in prosody modification (Rao 2005).

Since the LP residual is used for incorporating the desired prosody modification, there is no significant distortion due to resampling the residual samples both in the voiced and in the nonvoiced regions. This is because there is less correlation among samples in the LP residual compared to the correlation among the signal samples.

5. Practical issues in developing speech system using epoch-based analysis

The epoch-based analysis methods and applications discussed in this paper depend critically on our ability to extract epochs from speech signals. Epoch extraction using the ZFR output is robust against degradation caused by additive noise. However, in practice, the speech signal is degraded due to several factors such as reverberation in distance speech, channel effects in

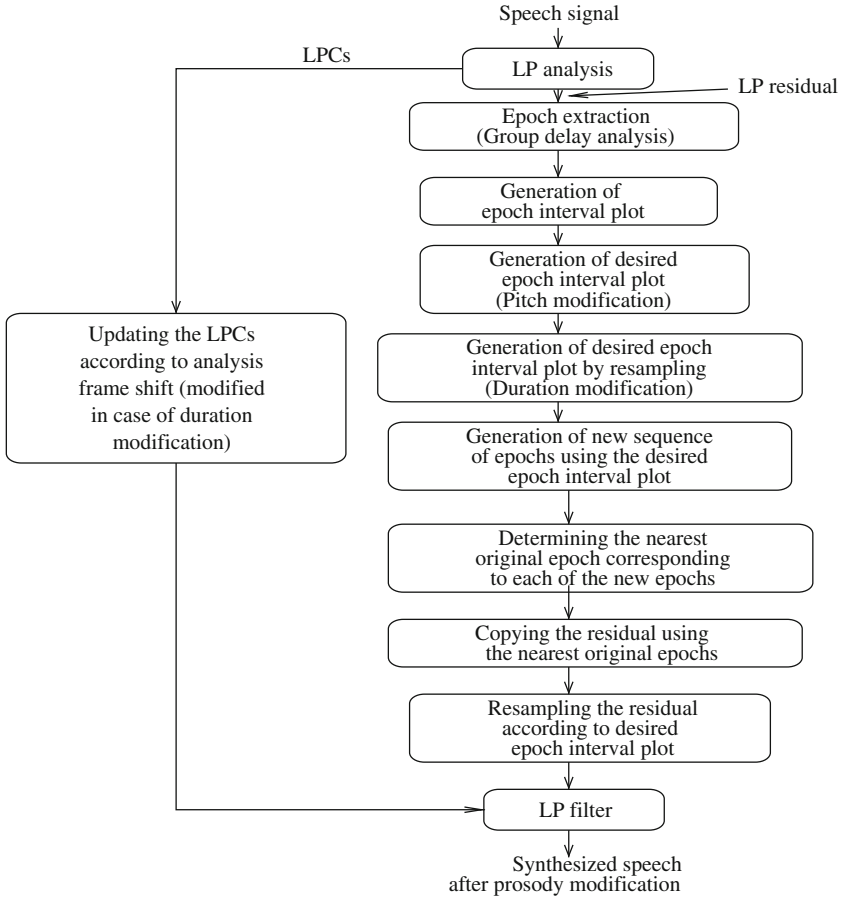

Figure 28. Block diagram for prosody modification (Rao 2005).

communication, recording devices, source coding methods used in communication and effects due to speech from multiple speakers.

Some of the effects of degradation can be overcome by deriving the epochs from the HE of the LP residual of the speech signal (Yegnanarayana *et al* 2011). In most such cases, the peaks of the HE will be stronger and impulse-like around the epochs. Hence, the epochs can be derived by using the ZFF on the HE of the LP residual. However, if there are too many spurious impulses in the signal caused by echoes or reverberation, then it is not possible to derive the epochs from such signals. In such cases, the effects of spurious impulses can be reduced by processing the data collected from number of spatially distributed microphones. By coherently adding the HE of the LP residual, after compensation for the time-delay between a pair of microphones, it may be possible to enhance the strength of the peaks in HE due to direct speech at the microphone locations. However, even then it may be difficult to extract all the epochs from the degraded signals. The effectiveness of human listening may tempt us to believe that it may be possible to extract all the epochs. However, human listening uses selective attention in perceiving speech, and hence all epochs may not be used in perceived pitch and other phonetic information. In fact, the redundancy in speech at various levels makes it possible to get the message in the signal,

even if some important events are missing due to degradations. Hence, determining the relevant information at all levels, including at the epochs level is a major challenge in speech signal processing.

Processing multispeaker mixed data to extract epochs for each speaker is a highly challenging task, as it involves separation of speech due to individual speakers, and then extracting the epochs for each speaker's speech. One way of dealing with this problem is by processing the mixed data collected at several spatially distributed microphones. As of now, the success of this approach is limited by the fact that it is difficult to extract the HE of the LP residual of individual speakers by compensating for time delay of the direct speech of the speaker at a pair of microphones. Here again the human listening may be adopting selective attention at various levels to get the message of each speaker from the mixed signals.

Most communication situations employ some form of coding the speech signal using source coding approach. In these cases, the waveform information is sought to be preserved as much as possible in some mean squared sense. It is likely that in such cases the epoch information may be modified significantly, and hence it may not be possible to derive the epoch information from the received signal. It is likely that perceptually significant component of source information may have been lost in the coding process. This may also suggest the need for exploring source coding methods that preserve the significant component of the excitation source information.

Material for this tutorial paper is borrowed from several theses (MS & PhD) and papers written in the Speech and Vision Laboratory at IIT Madras, Chennai and IIIT Hyderabad over the past several years. The authors would like to thank their colleagues and students who have contributed to the work presented in this paper over the recent decade or so.

References

Abberton E R M, Howard D M, Fourcin A J 1989 Laryngographic assessment of normal voice: A tutorial. *Clinical Linguistics and Phonetics* 3(3): 263–296

Ambramson A S, Lisker L 1965 Voice onset time in stop consonants: acoustic analysis and synthesis. *Proc. 5th Int. Congr. Phonetic Sciences*, Liege, A51

Ananthapadmanabha T V, Fant G 1982 Calculations of true glottal volume-velocity and its components. *Speech Commun.* 1: 167–184

Ananthapadmanabha T V, Yegnanarayana B 1975 Epoch extraction of voiced speech. *IEEE Trans. Acoust. Speech Signal Process.* 23(6): 562–570

Ananthapadmanabha T V, Yegnanarayana B 1979 Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Trans. Acoust. Speech Signal Process.* 27(4): 309–319

Atal B S, Hanauer S L 1971 Speech analysis and synthesis by linear prediction of the speech wave. *J. Acoust. Soc. Am.* 50(2): 637–655

Bachorowski J, Smoski M, Owren M 2001 The acoustic features of human laughter. *J. Acoust. Soc. Am.* 111: 1582–1597

Bagshaw P C, Hiller S M, Jack M A 1993 Enhanced pitch tracking and the processing of F_0 contours for computer and intonation teaching. *Proc. European Conf. on Speech Commun. (Eurospeech)*, Berlin, Germany, 1003–1006. URL http://www.cstr.ed.ac.uk/research/projects/fda/

Bapineedu G 2010 *Analysis of Lombard effect speech and its application in speaker verification for imposter detection*. MS thesis, Language Technologies Research Centre, International Institute of Information Technology, Hyderabad, India

Barros A K, Rutkowski T, Itakura F, Ohnishi N 2002 Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavlets. *IEEE Trans. Neural Netw.* 13: 888–893

Boersma P 1993 Accurate short-term analysis of fundamental frequency and the hormincs-to-noise ratio of a sampled sound. *Proc. Inst. Phonetic Sci.* 17: 97–110

Boll S F 1979 Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.* ASSP-27: 113–120

Borroff M L 2007 *A landmark underspecification account of the patterning of glottal stop*. PhD thesis, Stony Brook University, New York

Brookes M, Naylor P A, Gudnason J 2006 A quantitative assessment of group delay methods for identifying glottal closures in voiced speech. *IEEE Trans. Audio Speech Lang. Process.* 14(2): 456–466

Cardoso J-F 1998 Blind signal separation: statistical principles. *Proc. IEEE* 86: 2009–2025

Cheng Y M, O'Shaughnessy D 1989 Automatic and reliable estimation of glottal closure instant and period. *IEEE Trans. Acoust. Speech Signal Process.* 27: 1805–1815

Cheng Y M, O'Shaughnessy D 1991 Speech enhancement based conceptually on auditory evidence. *IEEE Trans. Signal Process.* 39: 1943–1954

CMU-ARCTIC speech synthesis databases. URL http://festvox.org/cmu_arctic/index.html

d'Alessandro C, Scherer K R 2003 Voice quality: Functions, analysis and synthesis (VOQUAL'03). ISCA Tutorial and Research Workshop, Geneva, Switzerland, http://archives.limsi.fr/VOQUAL/voicematerial.html (last viewed 04/08/2009)

de Cheveigne A, Kawahara H 2002 YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.* 111(4): 1917–1930

Dilley L, Shattuck-Hufnagel S, Ostendorf M 1996 Glottalization of word-initial vowels as a function of prosodic structure. *J. Phonetics* 24: 423–444

Ephraim Y, Van Trees H L 1995 A signal subspace approach for speech enhancement. *IEEE Trans. Speech Audio Process.* 3(4): 251–266

FFabre P 1957 Un procede electrique percutane d'inscrition de l'accolement glottique au cours de la phonation: glottographie de haute frequence. premiers resultats. *Bull. Acad. Natl. Med.* 141: 66

Flanagan J L, Jonston J D, Zahn R, Elko G W 1985 Computer-steered microphone arrays for sound transduction in large rooms. *J. Acoust. Soc. Am.* 78(5): 1508–1518

Fourcin A J, Abberton E 1971 First applications of a new laryngograph. *Med. Biol. Illus.* 21: 172–182

Frazier R H *et al* 1990 Enhancement of speech by adaptive filtering. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* New York, NY, USA

Frokjaer-Jensen B 1967 A photo-electric glottograph. *Annual Report of the Institute of Phonetics of University of Copenhagen* 2: 5–19

Frokjaer-Jensen B, Thorvaldsen P 1968 Construction of a fabre glottograph. *ARIPUC* 3: 1

Gauffin J, Sundberg J 1989 Spectral correlates of glottal voice source waveform characteristics. *J. Speech. Hear. Res.* 32: 556–565

Goldberg R, Riek L 2000 *A practical handbook of speech coders*. (Boca Raton, FL: CRC Press)

Gordon M, Ladefoged P 2001 Phonation types: a cross-linguistic overview. *J. Phonetics* 29(4): 383–406

Guruprasad S 2010 *Significance of processing regions of high signal-to-noise ratio in speech signals*. PhD thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras

Guruprasad S, Yegnanarayana B 2009 Perceived loudness of speech based on the characteristics of excitation source. *J. Acoust. Soc. Am.* 126(4): 2061–2071

Guruprasad S, Yegnanarayana B 2011 Performance of an event-based instantaneous fundamental frequency estimator for distant speech signals. *IEEE Trans. Audio Speech Lang. Process.* 19(7): 1853–1864

Hamon C, Moulines E, Charpentier F 1989 A diphone synthesis system based on time domain prosodic modifications of speech. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* Glasgow, 238–241

Hermes D J 1988 Measurement of pitch by subharmonic summation. *J. Acoust. Soc. Am.* 83(1): 257–264

Hess W, Indefrey H 1987 Accurate time-domain pitch determination of speech signals by means of a laryngograph. *Speech Commun.* 6: 55–68

Huang J, Zhao Y 1998 An energy-constrained signal subspace method for speech enhancement and recognition in colored noise. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* Seattle, WA, USA, 377–380

Huckvale M 2000 Speech filing system: Tools for speech research. URL http://www.phon.ucl.ac.uk/resource/sfs/

Iskra D, Grosskopf B, Marasek K, Van Den Heuvel H, Diehl F, Kiessling A 2002 SPEECON – speech databases for consumer devices: Database specification and validation. *Proc. Third Int. Conf. Lang. Resources Eval. (LREC)*, Las Palmas, Canary Islands - Spain, 329–333

Jankowski C R Jr, Quatieri T F, Reynolds D A 1995 Measuring fine structure in speech: Application to speaker identification. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* Detroit, MI, USA, 325–328

Jensen J, Hansen J H L 2001 Speech enhancement using a constrained iterative sinusoidal model. *IEEE Trans. Speech Audio Process.* 9(7): 731–740

Joseph A M, Guruprasad S, Yegnanarayana B 2006 Extracting formants from short segments using group delay functions. *Proc. Int. Conf. Spoken Language Processing*, Pittsburgh, USA, 1009–1012

Joseph A M, Yegnanarayana B, Gupta S, Kesheorey M R 2009 Speaker dependent mapping for low bit rate coding of throat microphone speech. *Proc. Interspeech 2009*, Brighton, UK, 1087–1090

Kominek J, Black A 2004 The CMU Arctic speech databases. *Proc. 5th ISCA Speech Synthesis Workshop*, Pittsburgh, USA, 223–224

Kounoudes A, Naylor P A, Brookes M 2002 The DYPSA algorithm for estimation of glottal closure instants in voiced speech. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* vol. 11, Orlando, FL, 349–352

Kumar S K 2010 *Analysis of laugh signals for automatic detection and synthesis*. MS thesis, Language Technologies Research Centre, International Institute of Information Technology, Hyderabad, India

Lecluse F L E 1977 *Elecroglottography*. Dissertation, Univ. of Rotterdam

Lee C K, Childers D G 1988 Cochannel speech separation. *J. Acoust. Soc. Am.* 83(1): 274–280

Leslau W 1995 *Reference grammar of Amharic*. (Wiesbaden: Otto Harrassowitz)

Lombard E 1911 Le signe de l'elevation de la voix, annals maladiers oreille. *Larynx. Nez. Pharynx.* 37: 101–119

Ma Y K C, Willems L F 1994 A Frobenius norm approach to glottal closure detection from the speech signal. *IEEE Trans. Speech Audio Process.* 2: 258–265

Makhoul J 1975 Linear prediction: A tutorial review. *Proc. IEEE*, 63(4): 561–580

Martin A, Doddington G, Kamm T, Ordowski M, Przybocki M 1977 The DET curve in assessment of detection task performance. *Proc. European Conf. Speech Process. Technol.* Greece, 1895–1898

McKenna J G 2001 Automatic glottal closed-phase location and analysis by Kalman filtering. *Proc. 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Perthshire, Scotland

Meyer G F 1995 Keele pitch database, School of Psychology, University of Liverpool, UK. URL http://www.liv.ac.uk/Psychology/hmp/projects/pitch.html

Mitchell O M M, Ross C A, Yates G H 1971 Signal processing for a cocktail party effect. *J. Acoust. Soc. Am.* 50: 656–660

Mittal U, Phamdo N 2000 Signal/noise klt based approach for enhancing speech degraded by colored noise. *IEEE Trans. Speech Audio Process.* 8: 159–167

Miyoshi M, Kaneda Y 1988 Inverse filtering of room acoustics. *IEEE Trans. Acoust. Speech Signal Process.* ASSP-36: 145–152

Monsen R B, Engebretson A M 1977 Study of variations in the male and female glottal wave. *J. Acoust. Soc. Am.* 62(4): 981–993

Morgan D P, George E B, Lee L T, Kay S M 1997 Cochannel speech separation by harmonic enhancement and supression. *IEEE Trans. Speech Audio Process.* 5: 407–424

Murty K S R 2009 *Significance of Excitation Source Information for Speech Analysis*. PhD thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras

Murty K S R, Yegnanarayana B 2006 Combining evidence from residual phase and MFCC features for speaker recognition. *IEEE Signal Process. Lett.* 13(1): 52–56

Murty K S R, Yegnanarayana B 2008 Epoch extraction from speech signals. *IEEE Trans. Audio Speech Lang. Process.* 16(8): 1602–1613

Murty K S R, Khurana S, Itankar Y U, Kesheorey M R, Yegnanarayana B 2008 Efficient representation of throat microphone speech. *Proc. Interspeech 2008*, Brisbane, Australia, 2610–2613

Murty K S R, Yegnanarayana B, Joseph M A 2009 Characterization of glottal activity from speech signals. *IEEE Signal Process. Lett.* 16(6): 469–472

Murty P S, Yegnanarayana B 1999 Robustness of group-delay-based method for extraction of significant excitation from speech signals. *IEEE Trans. Speech Audio Process.* 7(6): 609–619

Navarro-Mesa J L, Lleida-Solano E, Moreno-Bilbao A 2001 A new method for epoch detection based on the Cohen's class of time frequency representations. *IEEE Signal Process. Lett.* 8(8): 225–227

Naylor P A, Kounoudes A, Gudnason J, Brookes M 2007 Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. *IEEE Trans. Audio Speech Lang. Process.* 15(1): 34–43

Nemer E, Goubran R, Mahmoud S 2002 Speech enhancement using fourth-order cumulants and optimum filters in the subband domain. *Speech Commun.* 36: 219–246

Neocleous A, Naylor P A 1998 Voice source parameters for speaker verification. *Proc. Eur. Signal Process. Conf.* 697–700

Noisex-92. URL http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

Oh S, Viswanathan V 1992 Hands-free voice communication in an automobile with a microphone array. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* San Francisco, California, 281–284

Parsons T W 1976 Separation of speech from interfering speech by means of harmonic selection. *J. Acoust. Soc. Am.* 60: 911–918

Plante F, Meyer G F, Ainsworth W A 1995 A pitch extraction reference database. *Proc. European Conf. Speech Commun. (Eurospeech)* Madrid, Spain, 827–840

Plumpe M D, Quatieri T F, Reynolds D A 1999 Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. Acoust. Speech Signal Process.* 7: 569–586

Prasanna S R M 2004 *Event Based Analysis of Speech*. PhD thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras

Rao K S 2005 *Acquisition and incorporation of prosody knowledge for speech systems in Indian languages*. PhD thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras

Rao K S, Yegnanarayana B 2006 Prosody modification using instants of significant excitation. *IEEE Trans. Audio, Speech Lang. Process.* 14(3): 972–980

Rao K S, Prasanna S R M, Yegnanarayana B 2007 Determination of instants of significant excitation in speech using Hilbert envelope and group-delay function. *IEEE Signal Process. Lett.* 14(10): 762–765

Reddy S H M 2010 *Analysis of Speech at Different Speaking Rates using Excitation Source Information*. MS thesis, Language Technologies Research Centre, International Institute of Information Technology, Hyderabad, India

Reddy S H M, Prahallad K, Gangashetty S V, Yegnanarayana B 2010 Significance of pitch synchronous analysis for speaker recognition using AANN models. *Proc. Interspeech 2010*, Makuhari, Chiba, Japan, 669–672

Swamy R K, Murty K S R, Yegnanarayana B 2007 Determining number of speakers from multispeaker speech signals using excitation source information. *IEEE Signal Process. Lett.* 14(7): 481–484

Satyanarayana P 1999 *Short segment analysis of speech for enhancement*. PhD thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg., Chennai, India

Scalart P, Benmar A 1996 A system for speech enhancement in the context of hands-free radiotelephony with combined noise reduction and acoustic echo cancellation. *Speech Commun.* 20: 203–214

Scherer R C, Druker D G, Titze I R 1988 *Vocal physiology: Voice production mechanisms and functions*. (New York: Raven Press Ltd.)

Silverman H F 1987 Some analysis of microphone arrays for speech data acquisition. *IEEE Trans. Acoust. Speech Signal Process.* ASSP-35(12): 1699–1712

Smits R, Yegnanarayana B 1995 Determination of instants of significant excitation in speech using group delay function. *IEEE Trans. Speech Audio Process.* 3(5): 325–333

Sobakin A N 1972 Digital computer determination of formant parameters of the vocal tract from a speech signal. *Soviet Phys.-Acoust.* 18: 84–90

Stevens K N 1977 Physics of laryngeal behavior and larynx models. *Phonetica* 34: 264–279

Stevens M, Hajek J 2004 A preliminary investigation of some acoustic characteristics of ejectives in Waima'a: VOT and closure duration. In S Cassidy, F Cox, R Mannell, S Palethorpe (eds) *Proc. Tenth Australian Int. Conf. Speech Science and Technology*, Macquarie University, Sydney, ASSTA, 277–282

Strube H W 1974 Determination of the instant of glottal closures from the speech wave. *J. Acoust. Soc. Am.* 56: 1625–1629

Subramaniam S, Petropulu A P, Wendt C 1996 Cepstrum-based deconvolution for speech dereverberation. *IEEE Trans. Speech Audio Process.* 4: 392–396

Sun X 2002 Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. *Proc. IEEE Int. Conf. Acoust. Speech Signal Processing*, Orlando, FL, USA, 333–336

Talkin D 1995 *A robust algorithm for pitch tracking (RAPT)*. Speech coding and synthesis, (Amsterdam: Elsevier Science)

Tuan V N, d'Alessandro C 1999 Robust glottal closure detection using the wavelet transform. *Proc. European Conf. Speech Processing, Technology*, Budapest, 2805–2808

Van Den Berg J 1958 Myoelastic-aerodynamic theory of voice production. *J. Speech Hearing* 1: 227–244

Varga A, Steeneken H J M 1993 Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Comm.* 12(3): 247–251. [Online] Available: http://www.speech.cs.smu.edu/comp.speech/Section1/Data/noisex.html

Vassiére J 1997 Phonological use of the larynx. In ISCA LARYNX-1997, Marseille, France, 115–126

Veeneman D, BeMent S 1985 Automatic glottal inverse filtering from speech and electroglottographic signals. *IEEE Trans. Signal Process.* 33: 369–377

Wallace C 2007 The phonetics of laughter – A linguistic approach. *Interdisciplinary Workshop on the Phonetics of Laughter*, Saarbrucken, August 4–5 Saarbruchen

Wong D Y, Markel J D, Gray A H 1979 Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Trans. Acoust. Speech Signal Process.* 27: 350–355

Worku H S 2010 *Acoustic characterization of glottal stop and glottalized sounds in Amharic using non-spectral methods of speech analysis*. PhD thesis, Language Technologies Research Centre, International Institute of Information Technology, Hyderabad, India

Yegnanarayana B 1978 Formant extraction from linear prediction phase spectra. *J. Acoust. Soc. Am.* 63(5): 1638–1640

Yegnanarayana B, Murty P S 2000 Enhancement of reverberant speech using LP residual signal. *IEEE Trans. Speech Audio Process.* 8(3): 267–281

Yegnanarayana B, Smits R L H M 1995 A robust method for determining instants of major excitations in voiced speech. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Detroit, USA, 776–779

Yegnanarayana B, Veldhuis R N J 1998 Extraction of vocal-tract system characteristics from speech signals. *IEEE Trans. Speech and Audio Process.* 6(4): 313–327

Yegnanarayana B, Avendaño C, Hermansky H, Murty P S 1997 Processing linear prediction residual for speech enhancement. *Proc. European Conf. Speech Process. Technol.* Rhodes, Greece, 1399–1402

Yegnanarayana B, Avendaño C, Hermansky H, Murthy P S 1999 Speech enhancement using linear prediction residual. *Speech Commun.* 28(1): 25–42

Yegnanarayana B, Prasanna S R M, Duraiswami R, Zotkin D 2005 Processing of reverberent speech for time-delay estimation. *IEEE Trans. Speech Audio Process.* 13(6): 1110–1118

Yegnanarayana B, Murty K S R, Rajendran S 2008 Analysis of stop consonants in indian languages using excitation source information in speech signal. *Proc. Workshop Speech Anal. Process. Knowledge Discovery*, June 4–6 Aalborg, Denmark

Yegnanarayana B, Murty K S R 2009 Event-based instantaneous fundamental frequency estimation from speech signals. *IEEE Trans. Audio Speech Lang. Process.* 17(4): 614–624

Yegnanarayana B, Prasanna S R M, Guruprasad S 2011 Study of robustness of zero frequency resonator method for extraction of fundamental frequency. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* Prague, Czech Republic