Characterization of Glottal Activity From Speech Signals

K. Sri Rama Murty, B. Yegnanarayana, *Senior Member, IEEE*, and M. Anand Joseph

***Abstract*—The objective of this work is to characterize certain important features of excitation of speech, namely, detecting the regions of glottal activity and estimating the strength of excitation in each glottal cycle. The proposed method is based on the assumption that the excitation to the vocal-tract system can be approximated by a sequence of impulses of varying strengths. The effect due to an impulse in the time-domain is spread uniformly across the frequency-domain including at zero-frequency. We propose the use of a zero-frequency resonator to extract the characteristics of excitation source from speech signals by filtering out most of the time-varying vocal-tract information. The regions of glottal activity and the strengths of excitation estimated from the speech signal are in close agreement with those observed from the simultaneously recorded electro-glotto-graph signals. The performance of the proposed glottal activity detection is evaluated under different noisy environments at varying levels of degradation.**

***Index Terms*—Glottal activity detection, strength of excitation, zero-frequency resonator.**

I. Introduction

THE primary mode of excitation of the vocal-tract system during speech production is due to vibration of vocal folds (glottal activity). The strength of excitation during the glottal activity is determined mostly by the rate of closure of the vocal folds in each glottal cycle [1]. Detecting the regions of glottal activity and the strength of excitation in each glottal cycle from speech signal is a challenging task, as it is difficult to suppress the response of the time-varying vocal-tract system in the speech signal. Several methods have been suggested in the literature, which involve estimating the characteristics of the time-varying vocal-tract system, and then performing some form of inverse filtering of speech to highlight the characteristics of the excitation source [2]–[4]. Linear prediction (LP) analysis is one such method in which the LP coefficients are used to inverse filter the speech signal to derive the LP residual [5]. The LP residual has noise-like characteristics in the regions of nonglottal activity. In the regions of glottal activity, corresponding to the vocal fold vibration, the LP residual shows regions of large and small energies. The large energy region corresponds mostly to the closing phase of each glottal cycle. The effectiveness of detecting glottal activity from the LP residual depends on the accuracy of the LP model, and also the nature and quality of the speech signal.

In this paper, we propose a method based on the zero-frequency filtered signal to detect the regions of glottal activity and to estimate the strength of excitation in each glottal cycle. In Section II, we discuss the use of the zero-frequency resonator for highlighting the excitation source information in speech signals by filtering out the time-varying vocal-tract information. In Section III, we present a method to estimate the strength of excitation at epoch locations from the speech signals. Section IV discusses a method to automatically detect the regions of glottal activity, and evaluate its performance. In Section V we summarize the contributions of this paper.

Manuscript received November 03, 2008; revised January 26, 2009. Current version published April 24, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Philip C. Loizou.

K. Sri Rama Murty is with Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India (e-mail: ksrmurty@gmail.com).

B. Yegnanarayana and M. Anand Joseph are with International Institute of Information Technology, Hyderabad 500 032, India (e-mail: yegna@iiit.ac.in; anandjm@research.iiit.ac.in).

Digital Object Identifier 10.1109/LSP.2009.2016829

II. Zero-Frequency Resonator for Extracting Excitation Source Information

During the production of voiced speech, the excitation to the vocal-tract system can be approximated by a sequence of impulses of varying strengths. The effect of discontinuity due to the impulse-like excitation spreads uniformly across the frequency range including at the zero-frequency [6], [7]. In other words, even the output of a zero-frequency filter should reflect information about the discontinuities caused by the impulse-like excitation. The advantage of choosing a zero-frequency filter is that the output is not affected by the characteristics of the vocal-tract system which has resonances at much higher frequencies.

In this work, an ideal zero-frequency resonator is used to filter the speech signal. An ideal zero-frequency resonator is a 2nd order infinite impulse response (IIR) filter with a pair of real poles on the unit circle. We propose the use of a cascade of two ideal zero-frequency resonators to characterize the discontinuities due to impulse-like excitation in voiced speech. A cascade of two ideal zero-frequency resonators provides a roll-off of 24 dB per octave, which effectively dampens all the high frequency components beyond zero-frequency. Filtering a speech signal twice through a zero-frequency resonator results in an output that grows/decays as a polynomial function of time. Fig. 1(b) shows the output of filtering process for a segment of speech signal shown in Fig. 1(a). The effect of discontinuities due to impulse-like excitation is overridden by large DC offset that arises due to filtering at zero-frequency. The characteristics of discontinuities can be highlighted by subtracting the local mean computed over a small window. A window size of about one to two times the average pitch period is adequate for local mean subtraction. The resulting mean subtracted signal is shown in Fig. 1(c) for the filtered output shown in Fig. 1(b). The mean subtracted signal is called the *zero-frequency filtered signal* or

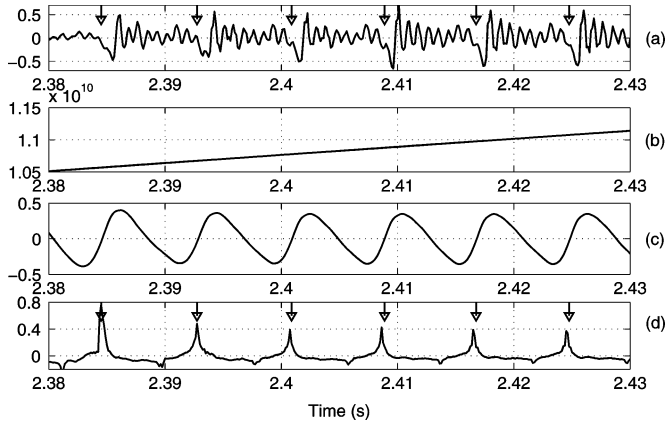

Fig. 1. Illustration of epoch extraction from speech signals. (a) A segment of speech signal taken from continuous speech. (b) Output of cascade of two ideal zero-frequency resonators. (c) Filtered signal obtained from mean subtraction. (d) DEGG signal. The arrows in (a) and (d) indicate the detected epoch locations.

merely the *filtered signal*. The following steps are involved in processing the speech signal to derive the filtered signal [6], [7]:

a) Difference the speech signal $s[n]$ to remove any DC component introduced by the recording device:

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

b) Pass the differenced speech signal $x[n]$ through a cascade of two ideal zero-frequency resonators. That is

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

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

c) Compute the average pitch period using the autocorrelation of 30 ms speech segments.

d) Remove the trend in $y_0[n]$ by subtracting the local mean computed at each sample. The resulting signal

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

is the zero-frequency filtered signal. Here $2N+1$ corresponds to the number of samples in the window used for mean subtraction. The choice of the window size is not critical as long as it is in the range of one to two pitch periods.

The filtered signal clearly shows sharper zero crossings around the epoch locations. In Fig. 1(c), the positive zero crossings are sharper than the negative zero crossings, and hence indicate the epoch locations. The locations of the positive zero crossings of the filtered signal in Fig. 1(c) coincide with the peaks in the differenced electro-glotto-graph (DEGG) signal shown in Fig. 1(d). The sharper zero crossings can either be positive zero crossings or negative zero crossings depending on the polarity of the signal (typically introduced by recording devices). The polarity of the sharper zero crossings can be automatically determined by comparing the slopes of the filtered signal around the positive zero crossings and the negative zero crossings over the entire duration of the utterance.

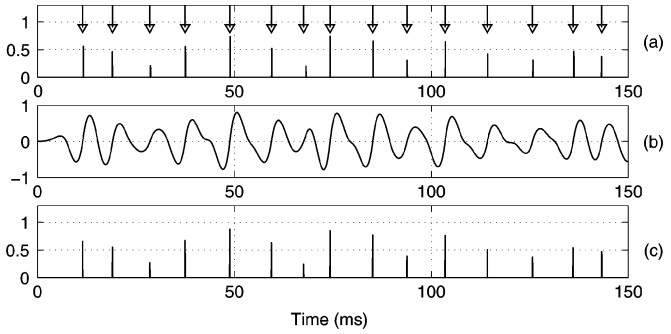

Fig. 2. (a) Sequence of randomly spaced impulses. (b) Zero frequency filtered signal. (c) Slope of signal around the positive (sharper) zero crossings. Arrows in (a) indicate hypothesized impulse locations.

III. Characterizing Strength of Excitation

The manner in which vocal folds vibrate influences the glottal airflow that serves as an excitation source for the vocal-tract filter. Sharper closure of the vocal folds corresponds to stronger excitation of the vocal-tract system. The peak intensity in the DEGG signal indicates the rate of glottal closure. Since the vocal-tract is known to absorb variable amount of acoustic energy, the acoustic pressure level as picked up by a microphone does not provide a reliable clue on the strength of excitation or the rate of glottal closure.

In this work, we exploit the narrowband nature of the zero-frequency resonator to measure the strength of excitation at each instant. Since the effect due to an impulse is spread uniformly across the frequency range, the relative strengths of impulses can be derived from a narrowband around any frequency, including the zero-frequency. Hence the information about the strength of excitation can also be derived from the zero-frequency resonator. It is observed that the slope of the zero-frequency filtered signal around the zero crossings corresponding to the epoch locations gives a measure of strength of excitation. Fig. 2(a) and (b) show a sequence of randomly spaced impulses with arbitrary strengths, and its zero-frequency filtered signal, respectively. The filtered signal [Fig. 2(b)] shows sharper zero crossings at the impulse locations, and the slopes of the filtered signal around those zero crossings are proportional to the actual impulse strengths as shown in Fig. 2(c).

This method of quantifying the epoch strength is valid even for speech signals. In the case of speech signals, the significant contribution at the zero-frequency is due to the impulse-like excitation. The vocal-tract system has resonances at much higher frequencies than zero-frequency. Hence the slope of the filtered signal around the epoch location predominantly reflects the strength of excitation. Fig. 3(d) shows the estimated strengths of excitation at the epoch locations for the speech signal shown in Fig. 3(a). Notice that the amplitude of the speech signal (Fig. 3(a)) around 0.5 s is low though the strength of the excitation as reflected in the DEGG signal (Fig. 3(b)) is high. The strength of excitation derived from the filtered signal of speech shows similar trend as that of the DEGG signal. Fig. 4(a) shows a scatter plot between the strength of excitation derived from the DEGG signal and the maximum absolute value of a sample of speech around the epoch location. Fig. 4(b) shows a scatter plot between the strength of DEGG

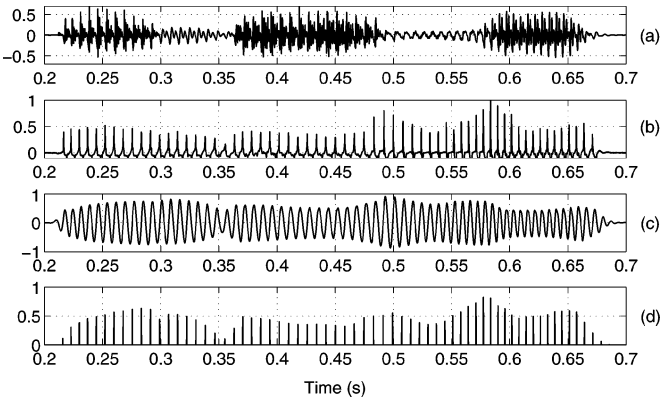

Fig. 3. (a) A segment of speech signal. (b) DEGG signal. (c) Filtered speech signal. (d) Slopes of the filtered signal around detected epoch locations (sharper zero crossings).

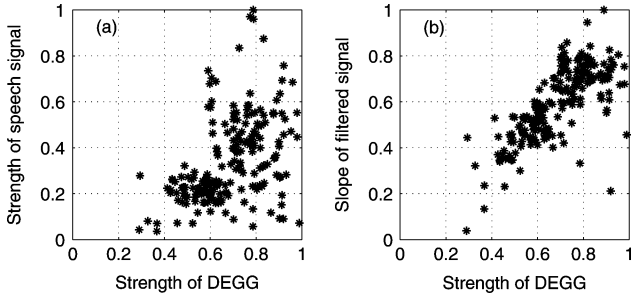

Fig. 4. Scatter plot of (a) DEGG versus speech signal and (b) DEGG versus slope of filtered signal.

signal and the strength of excitation estimated from the filtered signal of speech. The scatter plot in Fig. 4(b) shows a better linear orientation indicating that the estimated strength of excitation is proportional to the actual strength of excitation observed from EGG signal. This behavior is not present in Fig. 4(a), indicating that the strength of excitation can not be directly observed from the speech signal.

IV. Glottal Activity Detection

The strength of excitation of the vocal-tract system can be considered to be significant in the regions of the vocal fold vibration (glottal activity). In the absence of vocal fold vibration, the vocal-tract system can be considered to be excited by random noise, as in the case of fricatives. The energy of the random noise excitation is distributed both in time and frequency domains. While the energy of an impulse is distributed uniformly in the frequency domain, it is highly concentrated in the time-domain. As a result, the filtered signal exhibits significantly lower amplitude for random noise excitation compared to the impulse-like excitation. Hence the filtered signal can be used to detect the regions of glottal activity (vocal fold vibration) as illustrated in Fig. 5. Fig. 5(a) shows a segment of speech signal with regions of glottal activity, marked by dotted lines, obtained from the DEGG signal in Fig. 5(b). The filtered signal of speech shown in Fig. 5(c) clearly indicates the regions of glottal activity, and they match well with those obtained from the DEGG signal in

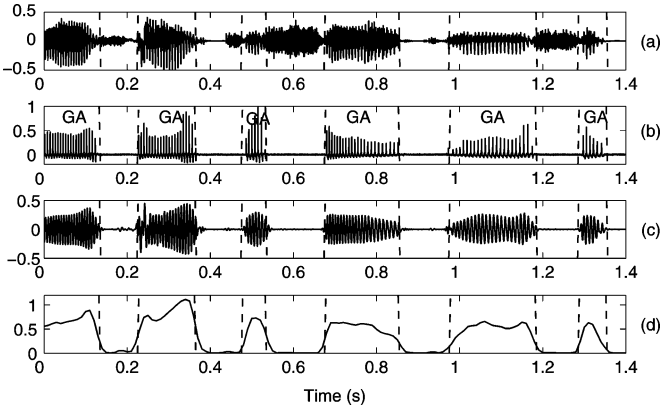

Fig. 5. Illustration of glottal activity detection from filtered signal. (a) Speech signal. (b) DEGG signal. (c) Filtered signal. (d) Energy of the filtered signal. Regions marked with GA in (b) indicate regions of glottal activity.

Fig. 5(b). Notice that the unvoiced regions around 0.6 s and 1.2 s in the speech signal (Fig. 5(a)) have very low amplitude in the filtered signal [Fig. 5(c)]. Hence the energy of the filtered signal shown in Fig. 5(d) can be used for glottal activity detection (GAD).

A. Performance Evaluation of Proposed GAD

The proposed GAD method was evaluated under different noisy environments at varying levels of degradation. A subset of CMU-Arctic database [8] consisting of 100 randomly selected sentences from each of the three speakers was used to evaluate the proposed GAD method. The entire dataset was samplewise labeled for glottal activity using the simultaneously recorded EGG signals available in the database. All the data was downsampled to 8 kHz.

To study the effect of noise on the proposed GAD, the method was evaluated on artificially generated noisy speech data. Several noise environments at varying levels of degradation were simulated by adding noise taken from Noisex-92 database [9]. The utterances were appended with silence such that in total amount of silence in each utterance is restricted to be about 60% of the data including pauses in the utterances. The database consists of speech signals under white, babble and vehicle noise environments at signal-to-noise ratio (SNR) ranging from 20 dB to 0 dB. The speech signals were processed using proposed zero-frequency resonator to obtain the filtered signal. The energy of the filtered signal for every frame of 20 ms at 100 frames/s is used to detect the glottal activity.

The performance of the proposed GAD method was evaluated using detection error tradeoff (DET) curves which show the tradeoff between false alarm rate (FAR) and false rejection rate (FRR). FAR represents the percentage of nonglottal activity frames that were detected as glottal activity, and FRR represents the percentage of glottal activity frames that were detected as nonglottal activity. 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 the accuracy of the GAD method. Fig. 6 shows the DET curves obtained for the proposed GAD algorithm under different noise environments at an SNR of 0 dB. The performance of GAD at varying

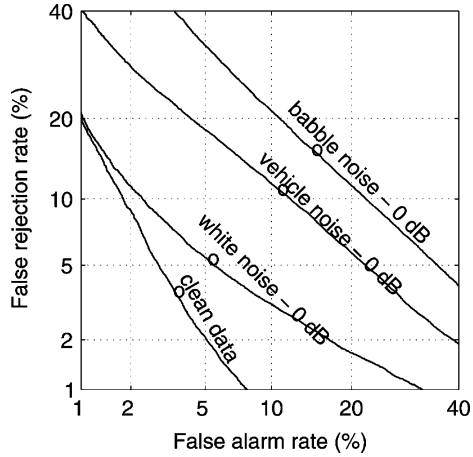

Fig. 6. DET curves indicating the performance of proposed GAD method under different noise environments.

TABLE I
PERFORMANCE OF GAD IN EER (%) UNDER DIFFERENT NOISE ENVIRONMENTS AT VARYING LEVELS OF DEGRADATION. REFERENCE IS DERIVED FROM EGG SIGNALS

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

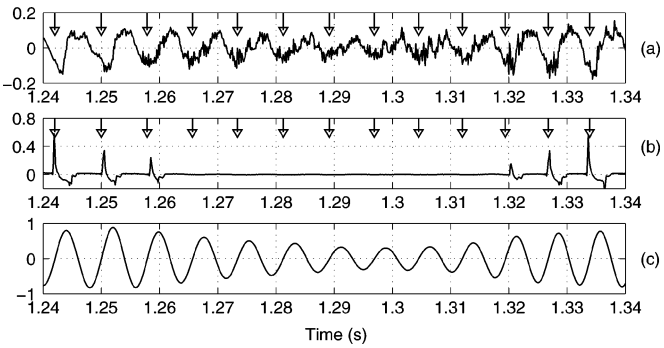

Fig. 7. Illustration of potential of proposed method in identifying weakly voiced regions. (a) A segment of speech signal. (b) DEGG signal. (c) Filtered signal of speech signal. Arrows in (a) and (b) indicate hypothesized epoch locations.

levels of degradation is listed in Table I using the reference derived from the EGG signals.

The proposed method achieved an EER of 3.54% on the clean data, and exhibits a gradual degradation under noisy conditions. The performance of the method under babble noise and vehicle noise is inferior to that under white noise because babble noise contains impulse-like excitations arising from epochs of other speakers, and vehicle noise introduces high degradations in low frequency components. The errors on clean speech may be attributed to the errors in the reference which are a result of inability of the EGG signals in capturing the weak voiced regions. Fig. 7(a) and (b) show a segment of weakly voiced region and its corresponding DEGG signal, respectively. The DEGG signal in Fig. 7(b) does not show prominent peaks around the epoch locations in the region from 1.26 s to 1.32 s, whereas the filtered signal in Fig. 7(c) clearly shows the glottal activity in that region, and the positive zero crossings approximately coincide with the epoch locations. Hence the proposed method can be effectively used to detect the glottal activity even in the weakly voiced regions. The performance of the proposed GAD under different noisy environments is evaluated with the reference derived from the clean speech. Table II gives the performance of the proposed GAD at varying levels of degradation using the reference derived from the clean speech data. The results show that the performance of the proposed method for GAD is robust against different types of degradation.

TABLE II
PERFORMANCE OF GAD IN EER (%) UNDER DIFFERENT NOISE ENVIRONMENTS AT VARYING LEVELS OF DEGRADATION. REFERENCE IS DERIVED FROM CLEAN SPEECH SIGNALS

Noise Type	20 dB	15 dB	10 dB	5 dB	0 dB
White	0	0	0.003	0.41	2.77
Babble	0	0.23	1.81	6.13	14.14
Vehicle	0	0.006	1.08	4.22	9.66

V. SUMMARY AND CONCLUSION

In this letter, we have proposed a method for detecting the regions of glottal activity and estimating the strength of excitation within each glottal cycle. The method exploits the impulse-like characteristic of the excitation which is extracted using a zero-frequency resonator. The method does not rely on estimating the vocal-tract response. The method is computationally very simple and also very accurate. The epoch location along with its strength of excitation form important features of a glottal pulse. This method may be useful in representing the excitation information in speech signal for speech coding and speech synthesis. The estimated strength of excitation may be useful in defining shimmer which is known to be a speaker-specific characteristic. The proposed method for GAD can be used to improve the existing voice activity detection methods.

REFERENCES

[1] P. Alku, T. Bakstrom, and E. Vikman, "Normalized amplitude quotient for parameterization of the glottal flow," *J. Acoust. Soc. Amer.*, vol. 112, no. 2, pp. 701–710, Aug. 2002.

[2] P. Alku, J. Vintturi, and E. Vilkman, "On the linearity of the relationship between the sound pressure level and the negative peak amplitude of the differentiated glottal flow in vowel production," *Speech Commun.*, vol. 28, pp. 269–281, Aug. 1999.

[3] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 325–333, Sep. 1995.

[4] Y. Stylianou, "Removing linear phase mismatches in concatenative speech synthesis," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 3, pp. 232–239, Mar. 2001.

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

[6] B. Yegnanarayana, K. S. R. Murty, and S. Rajendran, "Analysis of stop consonants in Indian languages using excitation source information in speech signal," in *Proc. ISCA-ITRW Workshop on Speech Analysis and Processing for Knowledge Discovery*, Aalborg, Denmark, Jun. 4–6, 2008.

[7] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[8] J. Kominek and A. Black, "The CMU arctic speech databases," in *Proc. 5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, 2004, pp. 223–224.

[9] *Noisex-92*, [Online]. Available: http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html