Voice Activity Detection in Degraded Speech Using Excitation Source Information

K. Sri Rama Murty[1], *B. Yegnanarayana*[2] *and S. Guruprasad*[1]

[1]Department of Computer Science and Engg., Indian Institute of Technology Madras, India
[2]International Institute of Information Technology Hyderabad, India
ksrm@cs.iitm.ernet.in, yegna@iiit.ac.in, guru@cs.iitm.ernet.in

Abstract

This paper proposes a method for detection of voiced regions from speech signals collected in noisy environment. The proposed method is based on the characteristics of excitation source of speech production. The degraded speech signal is processed by linear prediction analysis for deriving the linear prediction residual. Hilbert envelope of the linear prediction residual is processed using covariance analysis to obtain coherently-added covariance signal. The periodicity property of the coherently added covariance signal is exploited to detect the voiced regions using autocorrelation analysis. The performance of the proposed voice activity detection algorithm is evaluated under different noise environments and at different levels of degradation.

Index Terms: Excitation source information, linear prediction residual, glottal closure event, coherently added covariance signal.

1. Introduction

Speech signal can be considered as contiguous segments consisting of voiced, unvoiced and silence regions. The quality of speech collected in a noisy environment will be poor. But the voiced speech in the production process corresponds mostly to high Signal-to-Noise Ratio (SNR) regions, and hence these regions are less affected compared to nonvoiced and silence regions. Also, most (>80%) of the speech is of voiced type, and the voiced characteristics are retained even when speech is severely degraded. Therefore, if we are able to identify the voiced regions from the degraded signal, then such regions can be used for further processing in tasks like speech enhancement, speech coding and speaker recognition. The objective of the proposed Voice Activity Detection (VAD) algorithm is to identify voiced regions even when speech is degraded. The proposed algorithm may be used as a front-end processor for some of the applications mentioned above.

Conventional VAD algorithms assume that the background noise statistics are stationary over a longer period of time than those of speech, which makes it possible to estimate the time varying noise statistics in spite of occasional presence of speech. To determine the presence or absence of speech, the observed signal statistics in the current frame are compared with the estimated noise statistics according to some decision rules. Moreover, this initial decision is modified by a hang-over scheme to minimize misdetections at weak speech tails. In these methods, estimating the parameters governing the noise model is a crucial step. Sohn et al., proposed a decision-directed parameter estimation method and a HMM-based hang-over scheme which improves the performance of VAD [1]. In [2], Li et al., proposed a VAD scheme using the properties of higher order statistics of speech and noise signals. Their scheme employs the logarithm of kurtosis of the LP residual, and is shown to be more effective and efficient in detecting the speech activity in medium to low SNR conditions, without being affected by variations in the signal energy. A statistical method which employs a low-variance spectrum estimate, and determines an optimum threshold based on the estimated noise statistics is presented in [3]. An adaptive method of finding an appropriate statistical model for noisy speech in the spectral domain is presented in [4]. It is shown that the complex Laplacian and Gamma density functions are better suited for the parametric representation of noisy speech spectra distribution than the conventional Gaussian density function. Dependence of these methods on the statistical characterization of degradation constrains their usage to situations where such characteristics can be derived. Moreover, the statistical characteristics of degradations may vary widely depending on the type degradation. Hence, these methods may not be suitable to situations where the degradations are unknown and/or non-stationary. In this paper, we present a VAD algorithm that relies on the properties of the speech production process, rather than the statistical properties of the noise spectrum.

The speech-specific knowledge from vocal tract system or from excitation source or from both may be used for developing VAD algorithm. For example, the first predictor coefficient in LP analysis contains some discriminating information between voiced and non-voiced (unvoiced and silence) regions, and is used in VAD technique proposed in [5]. The vocal tract system features are severely affected by the degradations, compared to some features corresponding to the excitation source of speech production. For example, the relative spacing between the Glottal Closure (GC) events is not affected by degradations. In this paper we use this property of the excitation source of speech production for developing a VAD algorithm. The paper is organized as follows: Section 2 discusses extraction of the excitation source information from the speech signal. The proposed method for detection of voiced regions is discussed in Section 3. Performance evaluation of the proposed algorithm is presented in Section 4. Summary of the present work and scope for further studies in this direction are given in Section 5.

2. Excitation Source Information for VAD

The excitation source information can be extracted from the speech signal by performing LP analysis [6]. In LP analysis, the sample $s[n]$ is estimated as a linear weighted sum of past p

samples. The predicted sample $\hat{s}[n]$ is given by

$$\hat{s}[n] = -\sum_{k=1}^{p} a_k \, s[n-k], \tag{1}$$

where p is the order of prediction and $\{a_k\}$, $k = 1,\, 2, \ldots,\, p$, is the set of linear prediction coefficients (LPCs). The LPCs are obtained by minimizing the mean-squared error between the predicted sample value and actual sample value over the analysis frame. The error between the actual value $s[n]$ and the predicted value $\hat{s}[n]$ is given by

$$e[n] = s[n] - \hat{s}[n] = s[n] + \sum_{k=1}^{p} a_k s[n-k]. \tag{2}$$

The error $e[n]$ is called LP residual of the speech signal. Since $\{a_k\}$ models the vocal tract system features, the LP residual $e[n]$ contains mostly information about the excitation source. A segment of degraded speech signal and its LP residual are shown in Figs. 1(a) and (b), respectively. Whenever there is significant excitation to the vocal tract system, it is indicated by a large error in the LP residual. This can clearly be seen in the case of voiced speech, where the significant excitation within a pitch period coincides with the GC event. The GC event is the instant at which closure of the vocal folds takes place in each glottal cycle. Even though the LP residual contains mostly the excitation source information, there are difficulties in using it directly for further processing. This is due to fluctuations caused by the phase of the residual, which results in signal of random polarity around the instants of significant excitation. The effect of phase can be reduced by using amplitude envelope of the analytic signal derived from the LP residual [7]. The analytic signal $e_a[n]$ corresponding to the LP residual $e[n]$ is defined as [8]

$$e_a[n] = e[n] + je_h[n] \tag{3}$$

where $e_h[n]$ is the Hilbert transform of $e[n]$, and it is computed as

$$e_h[n] = IFT[E_h(\omega)]$$

where

$$E_h(\omega) = \begin{cases} jE(\omega), & -\pi \le \omega < 0 \\ -jE(\omega), & 0 \le \omega < \pi \end{cases} \tag{4}$$

Here IFT denotes the inverse Fourier transform, and $E(\omega)$ is the Fourier transform of $e[n]$. The amplitude envelope of the analytic signal $e_a[n]$ (also called Hilbert envelope $h_e[n]$) of the LP residual is given by

$$h_e[n] = |e_a[n]| = \sqrt{e^2[n] + e_h^2[n]} \tag{5}$$

The Hilbert Envelope (HE) of the LP residual is shown in Fig. 1(c). The instants of significant excitation show periodic nature in the voiced regions, and this periodicity is not present in non-voiced (noise and unvoiced) regions. The instants of significant excitation in the HE can further be emphasized using covariance analysis as described below.

Consider a frame of the HE $f_m[n]$ of N samples, given by

$$f_m[n] = h_e[n+m] \, w[n]$$

Here, $h_e[n]$ is the HE of the LP residual, m is the starting sample number of the frame and $w[n]$ is the window function of N samples given by

$$w[n] = \begin{cases} 1, & \text{for } 0 \le n < N \\ 0, & \text{otherwise.} \end{cases}$$

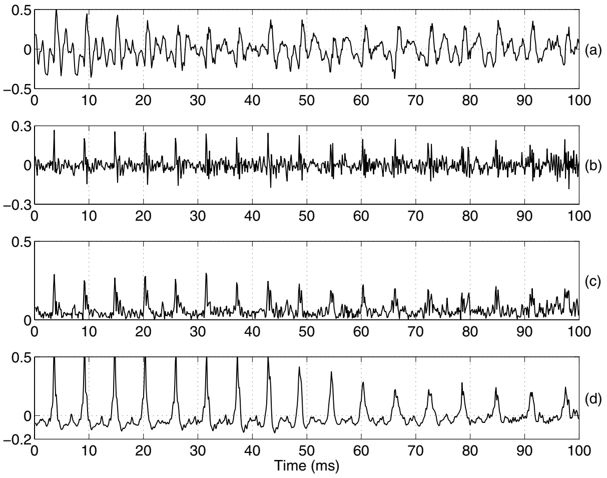

Figure 1: *(a) Segment of degraded speech signal and its (b) LP residual, (c) Hilbert envelope of the LP residual, (d) Coherently added covariance signal derived from the Hilbert envelope.*

For every frame $f_m[n]$ of N samples, the covariance sequence $\varphi_m[l]$ is computed as

$$\varphi_m[l] = \frac{\sum_{n=0}^{N-1} f_m[n] f_m[n+l]}{\sqrt{\sum_{n=0}^{N-1} f_m^2[n]} \sqrt{\sum_{n=0}^{N-1} f_m^2[n+l]}}, \qquad l = 1, 2, \ldots, N. \tag{6}$$

where l is the time-shift. The peaks in the covariance sequence occur at an interval corresponding to the peaks in the HE of the LP residual. The peaks in the covariance sequence can be time-aligned with the peaks in the current frame of the HE by using cross-correlation approach. The location of the strongest peak in the cross-correlation function of $f_m[n]$ and $\varphi_m[n]$ gives the delay k_m between the two sequences. The delay k_m is adjusted in the covariance sequence to align the peaks in the covariance sequence in coherence with the peaks in the current frame of the HE. The time-aligned covariance sequences are computed with a shift of q samples, and they are added to obtain the coherently added covariance signal $c[n]$ as

$$c[n] = \sum_{m} \varphi_m[n - m - k_m], \qquad m = 0,\ q,\ 2q, \ldots \tag{7}$$

For illustration, a segment of the HE of the LP residual for speech signal at 5 dB SNR is shown in Fig. 2(a). The time-aligned covariance sequences for four consecutive frames of size 20 ms with a frame shift of 2 ms of the HE are shown in Figs. 2(b)-(e). The time-aligned covariance sequences are coherently added to obtain the signal shown in Fig. 2(f). The peaks corresponding to the GC events are emphasized in the coherently-added covariance signal, as compared to the peaks in the HE. The periodic behavior of these peaks in the voiced regions of the coherently-added covariance signal can be exploited for developing a VAD algorithm. The normalized strength of the first peak (first major peak after the center peak) in the autocorrelation sequence of the coherently-added covariance signal is used to quantify the periodic behavior.

3. Voice Activity Detection Algorithm

To exploit the periodic behavior of peaks in coherently-added covariance signal of voiced regions, autocorrelation analysis is

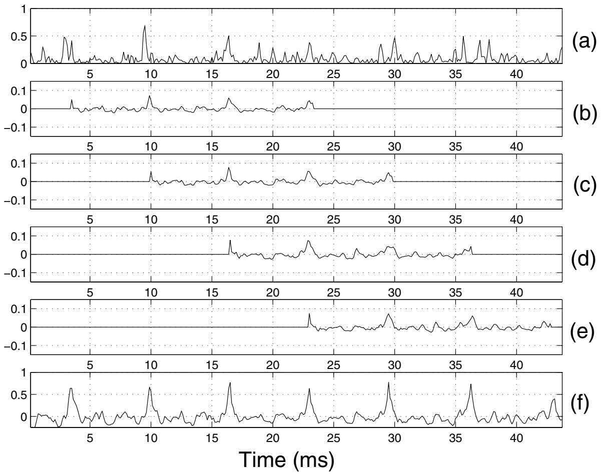

Figure 2: *(a) A segment of the Hilbert envelope of the LP residual, (b)-(e) time-aligned covariance sequences of four consecutive frames and (f) coherently-added covariance signal.*

performed. The strength of the first major peak (after the center peak) in the normalized autocorrelation sequence of the coherently added covariance signal is an indication of the periodicity and voicing level of the segment. For illustration, a 30 ms frame of coherently-added covariance signal of a voiced segment of degraded speech and its normalized autocorrelation sequence are shown in Figs. 3(a) and 3(b), respectively. Similarly, a 30 ms frame of non-voiced region and its normalized autocorrelation sequence are shown in Figs. 3(c) and 3(d), respectively. The values of the normalized strength P_s of the first peak are indicated in Figs. 3(b) and 3(d). The normalized strength of the first peak in the voiced regions is higher compared to that for the non-voiced regions. The normalized strength of the first peak is computed for every frame of the coherently-added covariance signal using a frame shift of one sample. The values of the peak strength thus obtained are used to detect the voice activity in the speech signal. A large value of peak strength indicates the presence of voicing in the corresponding frame.

4. Evaluation of VAD algorithm

The VAD algorithm was evaluated in different noisy environments at different levels of degradation. The data set required for evaluation was prepared as suggested in [3]. A subset of TIMIT corpus consisting of 62 individual speakers, each speaking 4 sentences, is used to evaluate the proposed VAD algorithm. This data consists of 248 different spoken sentences encompassing all phones and eight different dialects as defined in the TIMIT set. All data was downsampled to 8 kHz. Sentences were concatenated in sets of four, and silence was inserted between sentences. The duration of silence between sentences was randomly chosen; however, the duration of silence was constrained to 60% of the total duration. The resulting database consisted of 32 min of speech data, of which 40% was active speech, which is typically the amount of speech activity in a telephone conversation.

The entire data set was samplewise labeled for voice activity using the clean data. Several noise environments were artificially added to the clean data set at varying SNRs. The noise used was taken from the NOISEX-92 database consisted of babble, factory, Gaussian, high frequency, pink and vehicular noise. Including different noise environments and SNRs, the proposed VAD algorithm was evaluated on 15 hours of noisy speech.

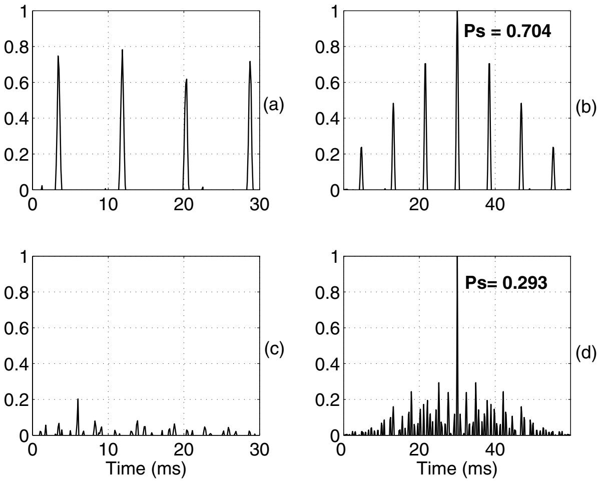

Figure 3: *(a) A 30 ms segment of the Hilbert envelope of voiced speech, and, (b) its autocorrelation sequence. (c) A 30 ms segment of the Hilbert envelope of nonvoiced speech, and (d) its autocorrelation sequence.*

A 10^{th} order LP analysis was performed on the degraded speech signal to obtain the LP residual. The HE of the LP residual was processed using covariance analysis to obtain coherently-added covariance signal. Autocorrelation analysis was performed on the coherently-added covariance signal using frames of size 25 ms with a shift of one sample. A segment of speech signal degraded by babble noise at 5 dB SNR, and corresponding actual voice active regions, are shown in Fig. 4(a) and Fig. 4(b), respectively. The sequence of peak strength values obtained from the autocorrelation analysis of coherently added covariance signal is shown in Fig. 4(c). The voice active regions are marked by high peak strength values, and hence the algorithm can be used for voice activity detection.

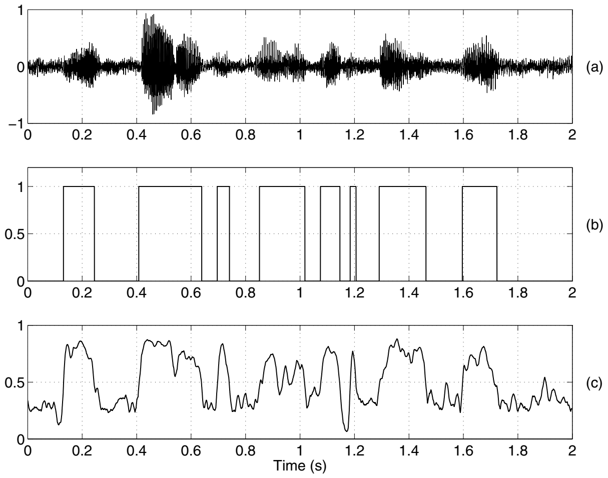

Figure 4: *(a) A segment of speech signal degraded by babble noise at 5 dB SNR (b) Manually marked voiced regions (c) Normalized peak strength values obtained from autocorrelation analysis.*

The performance of the proposed algorithm was evaluated using the Detection Error Tradeoff (DET) curves which show the tradeoff between False Alarm Rate (FAR) and False Rejection Rate (FRR). The FAR represents the number of non-voiced frames that were detected as voiced, whereas, the FRR represents the number of voiced frames that were not detected. The DET curves obtained by the proposed VAD algorithm under babble noise environment at different levels of degradation are given in Fig. 5. As the FAR decreases, the FRR increases and vice versa. 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 VAD algorithm. The performance of VAD in various noise environments at different degradation levels is listed in Table 1.

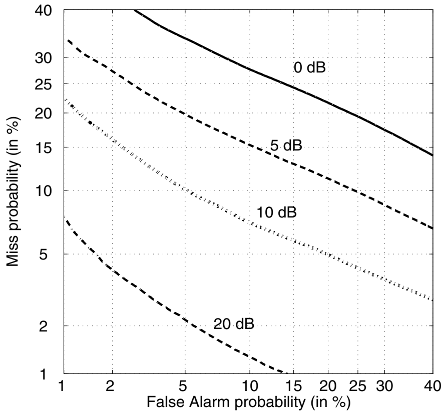

Figure 5: *DET curves indicating the performance of proposed VAD algorithm under babble noise environment at different levels of degradation.*

Table 1: *Performance of VAD in EER (%) for various noise environments and SNRs.*

Noise Type	20 dB	15 dB	10 dB	5 dB	0 dB
Babble	3	5	8.5	14	21
Factory	3.2	5.5	10	16	25
Gaussian	5	11	17	25	32
HF Channel	4	7	12	20	25
Pink	4	6	11	19	29
Vehicular	0	0	0.2	0.6	1.2

The average performance of the proposed excitation source based VAD algorithm is comparable with the existing spectrum based statistical methods [3] [4]. The proposed method is inferior to the spectrum based methods under white noise environment. This poor performance can be attributed to the limitations of LP analysis under high degradation due to white noise. The performance can be improved by characterizing the type of noise and performing a lower order LP analysis under white noise environment. The proposed VAD algorithm performs better than the spectrum based methods [3] under babble noise environment. The spectrum based methods perform poorly in babble noise environment because of its speech-like spectral properties. But, the excitation source information and the periodicity of the GC instants are not preserved in the babble noise, and hence the proposed algorithm is well suited for babble noise environment. The proposed method is comparable to the spectrum based methods in vehicular noise environment. Since the proposed method is based on the excitation source information, it provides complimentary information to the spectrum based methods. Moreover, it is difficult to develop a universal VAD algorithm that performs equally well in all the noise environments. Hence, the proposed method can be used along with the existing spectrum based methods to develop a robust VAD system.

5. Conclusions

In this paper, a method based on the characteristics of excitation source of speech production was proposed to detect the voiced regions in degraded speech. The method exploits the periodicity of the GC events in the excitation source information to identify the voiced regions. Because of high SNR nature of the regions around the GC events, the periodicity information is preserved even under high levels of degradation. Coherent addition of the covariance signals derived from HE of LP residual was proposed to emphasize the peaks at GC events. This method of emphasizing the peaks can also be employed to detect the accurate locations of GC events, which is essential for applications like speech enhancement [9]. The proposed method relies mainly on the characteristics of the speech production process, rather than on the properties of the noise spectrum. Moreover, the proposed method does not assume any noise characteristics, and does not depend on parameters estimated from the noise spectrum. Hence, the proposed method can be applied irrespective of the noise environment. The performance of the proposed method can be improved by employing a hang-over scheme which modify the VAD decision using sequence information.

6. References

[1] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Proc. Letters*, vol. 6, pp. 1–4, Jan. 1999.

[2] K. Li, M. N. S. Swamy, and M. O. Ahmad, “An improved voice activity detection using higher order statistics,” *IEEE Trans. Speech Audio Processing*, vol. 13, pp. 965–974, Sep. 2005.

[3] A. Davis, S. Nordholm, and R. Togneri, “Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold,” *IEEE Trans on Audio, Speech, and Language Processing*, vol. 14, pp. 412–424, Mar. 2006.

[4] J.-H. Chang, N. S. Kim, and S. K. Mitra, “Voice activity detection based on multiple statistical models,” *IEEE Trans. Signal Processing*, vol. 54, pp. 1965–1976, Jun. 2006.

[5] W. Hess, *Pitch determination of speech signals*. Berlin Heidelberg, New York: Springer-Verlag, 1983.

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

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

[8] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time signal processing*. Upper Saddle River, New Jersey: Prentice Hall, 2000.

[9] M. Chaitanya, *Single Channel Speech Enhancement*. M.S. Thesis, Indian Institute of Technology Madras, 2005.