Robust Voice Activity Detection using Group Delay Functions

Sree Hari Krishnan P, R.Padmanabhan, Hema A Murthy

Department of Computer Science and Engineering
Indian Institute of Technology Madras
Chennai 600036, India
{hari, padmanabhan, hema} @lantana.cs.iitm.ernet.in
http://www.lantana.tenet.res.in

Abstract—**In this paper, we present an algorithm for Voice Activity Detection (VAD) in speech signals with very low SNR. In the proposed algorithm, the short-term energy of the speech signal is viewed as the positive frequency part of the magnitude spectrum of a minimum phase signal. The group delay of this signal is then computed. The speech regions of the signal are characterized by well-defined peaks in the group delay spectrum, while the non-speech regions are identified by well-defined valleys. The proposed method is compared experimentally with G.729 Annex B in very low SNR conditions with different types of noises and is found to perform significantly better.**

I. Introduction

The purpose of a Voice Activity Detection (VAD) algorithm is to distinguish between speech and noise regions in a speech signal. It has applications in speech activity detection for automatic speech recognition (ASR), speech absence detection for noise estimation, speech coding and echo cancellation. In ASR, VAD improves recognition-accuracy significantly. While in speech coding, it aids in improving the system resource utilization in conjunction with comfort noise generator (CNG) during periods of speech absence by selectively encoding and transmitting data.

Early approaches to VAD extract features of a speech signal like short-term energy, pitch, short-term zero-crossing rate, etc and compare these with a threshold. Though these methods work well in high SNR environments, they fail in low SNR (<5dB) environments. Another difficulty with these thresholding methods is the selection of the threshold, owing to the variation amongst speakers and the presence of fricatives and nasals.

The G.729-Annex B VAD algorithm is currently recommended by International Telecommunication Union (ITU) [1]. It uses a piecewise linear discriminant based on line spectral frequencies, high and low band energies and zero-crossing rate to make the VAD-decision.

Statistical VAD techniques based on modeling speech and noise as independent Gaussian random variables were introduced in [2] and [3]. Modeling speech as a Laplacian random variable was proposed in [4]. These methods were reported to show better performance than G.729B. VAD for handling the non-stationarity of noise, was explored in [5] using a set of six fuzzy rules. Further methods which discuss the non-stationarity of noise were explored in [6], [7], [8].

More recently, some VAD algorithms exploit the fact that Higher Order Statistics (HOS) show promising results when dealing with a mixture of Gaussian and non-Gaussian processes [9], [10]. The application of HOS to speech processing has been motivated by the assumption that speech has certain HOS properties that are distinct from that of Gaussian noise. However, when noise is not Gaussian, or because some unvoiced speech behaves like Gaussian noise, the distinction between speech and noise using this method, in these cases, is not clear. To mitigate this, HOS combined with other parameters was proposed in [11].

In this paper, we describe an algorithm for VAD using the group delay function. Here, the short-term energy (STE) is treated as a magnitude spectrum of an arbitrary signal and is then converted to its minimum phase equivalent. The group delay of this equivalent signal is then obtained. The group delay resolves speech and noise well: while speech is characterized by peaks, noise is characterized by valleys. We demonstrate the robustness of this VAD algorithm, by applying it to speech with very low SNR (<5dB).

This paper is organized as follows: Section II reviews the group delay function. In Section III, the VAD algorithm using the minimum phase group delay function is presented. Comparison with G.729B in very low SNR conditions is performed in Section IV. Lastly, we draw some conclusions in Section V.

II. Review of the minimum phase group delay function

Let $x_i(n)$ and $X_i(\omega)$ be Fourier transform pairs, and

$$X_3(\omega) = X_1(\omega).X_2(\omega) \tag{1}$$

Then,

$$\mid X_3(\omega) \mid = \mid X_1(\omega) \mid . \mid X_2(\omega) \mid \tag{2}$$

$$arg(X_3(\omega)) = arg(X_1(\omega)) + arg(X_2(\omega)) \tag{3}$$

and

$$\tau_{x_3}(\omega) = \tau_{x_1}(\omega) + \tau_{x_2}(\omega) \tag{4}$$

where $\tau_{x_3}(\omega)$, $\tau_{x_1}(\omega)$ and $\tau_{x_2}(\omega)$ correspond to the group delay function of $X_3(\omega)$, $X_1(\omega)$, $X_2(\omega)$ respectively. From equations (1)-(3) we see that the multiplication in the spectral domain corresponds to addition in the group delay domain. The group delay derived from a minimum phase signal is called the *minimum phase group delay function.* From [12], we observe that the group delay function resolves the poles and zeros better than the magnitude and phase spectra if the signal is minimum phase. Further, it is also shown in [12] that a non-minimum phase signal can be converted into a minimum phase signal by taking the causal portion of the inverse Fourier transform of its magnitude spectrum. This is the feature that is exploited in next section for generating a minimum phase signal from the short-term energy function, which, in turn is used to develop a VAD algorithm.

III. VAD in low SNR using the minimum phase group delay function

Before describing the procedure for VAD using the minimum phase group delay function, the short-term energy with respect to a given speech signal is defined.

A. Short-term energy

The short-term energy (STE) of each frame of the speech signal is calculated using the formula given below:

$$E(m) = \sum_{n=0}^{N-1} (w(n)x(m-n))^2$$

where $w(n)$ is a window [13] and N is the frame-size in samples and m is a multiple of the frame-shift.

B. Algorithm for VAD

In [14], the group delay spectrum was used to extract syllable boundaries in clean environment, but in this paper, it is used to extract noise and speech regions in the signal in very low SNR conditions. We take the STE of the speech signal, view it as a magnitude spectrum, derive its minimum phase equivalent ($c(n)$) and finally compute its group delay function ($\tau(\omega)$). Next, for every peak in the group delay spectrum, the corresponding location of the valley before the peak and the location of the valley after the peak are identified. The region between these valleys correspond to a speech region. A formal description of this procedure is listed below.

1. Let $x(n)$ be the given speech signal.
2. Compute its STE. Let us denote it by $E(m)$ with $0 \leq m \leq K-1$ and K is the length of the STE defined as K = (length of speech signal in samples)/(frame-shift in samples).
3. Obtain $\tilde{E}(m)$ from $E(m)$ by zero padding, making $\tilde{E}(m)$ an exact power of two.

$$\tilde{E}(m) = E(m) \qquad 0 \leq m \leq K-1 \tag{5}$$

$$\tilde{E}(m) = 0 \qquad K \leq m \leq M-1 \tag{6}$$

where $M = 2^{\lceil \log_2(K) \rceil}$.

4. Form the symmetric-sequence $E_s(m)$

$$E_s(m) = \tilde{E}(m) \qquad 0 \leq m \leq M-1 \tag{7}$$

$$E_s(m) = \tilde{E}(2M-m-1) \qquad M < m \leq 2M-1 \tag{8}$$

where $2M$ is the DFT order as described next. This new sequence is considered as a magnitude spectrum of an arbitrary signal of $2M$ points between $-\pi$ and π and is denoted as $E_s(k)$. Let $2M = N$.

5. To reduce the dynamic range, perform the following:

$$\acute{E}_s(k) = E_s(k)^{\gamma} \qquad 0 \leq k \leq N-1, \qquad 0 < \gamma \leq 1$$

6. Compute the IDFT of the function $\acute{E}_s(k)$. The causal portion of the resulting sequence denoted by $e(n)$ is a minimum phase signal [15].
7. Compute the group delay function [12], [16] of $e(n)w(n)$, where $w(n)$ is a low-pass filter of length N_l. The group delay is computed as follows:
- Compute the phase spectrum $\phi(k)$ of $e(n)w(n)$.
- Compute the forward difference

$$\tau(k) = \phi(k) - \phi(k-1) \qquad 0 \leq n \leq N-1$$

where $\tau(k)$ is the group delay function.

8. For every peak i in the group delay function($\tau(k)$), compute the following:
- Identify the valley before the peak i as $f_b(i)$.
- Identify the valley after the peak i as $f_e(i)$.
- The frames between $f_b(i)$ and $f_e(i)$ denote speech regions.

9. Output the VAD sequence $V_{GD}(n)$ consisting of -1 and 1, where -1 denotes a noisy frame and 1 denotes a speech frame.

Window Scale Factor (WSF $= \frac{N}{N_l}$) and γ are used to control the resolution of the group delay. The sequence of steps is illustrated with an example in Figure 1.

IV. Results and discussion

A. Experimental setup

Experiments were performed on 432 speech files (216 female, 216 male) obtained by concatenating sets of three individual speech utterances taken from TIMIT [17]. These files were low-pass filtered and resampled to 8kHz to conform to G.729B. Six different types of noise (babble, HF channel, factory, volvo, pink and white) from the NOISEX-92 database were added in the range -15 dB to 15 dB to these files. Of these, results are tabulated below only for three types of noise for want of space.

B. The normalized correlation metric

To evaluate the effectiveness of the proposed group delay VAD algorithm (GD-VAD) and the G.729B VAD algorithm at various noise levels, we find out the similarity of each to a corresponding manually marked VAD sequence. To this end, the normalized cross-correlation at zeroth lag with the manually marked VAD sequence is used. This value has a maximum of 1 which indicates a total match of all frames with the manually marked VAD, and a minimum value of -1 which indicates a total mismatch. Formally, this is done as follows:

1. For every noisy speech signal $x(n)$, get the GD-marked, G.729B-marked and manually-marked VAD sequences, $V_{GD}(n)$, $V_{G729}(n)$ and $V_M(n)$.

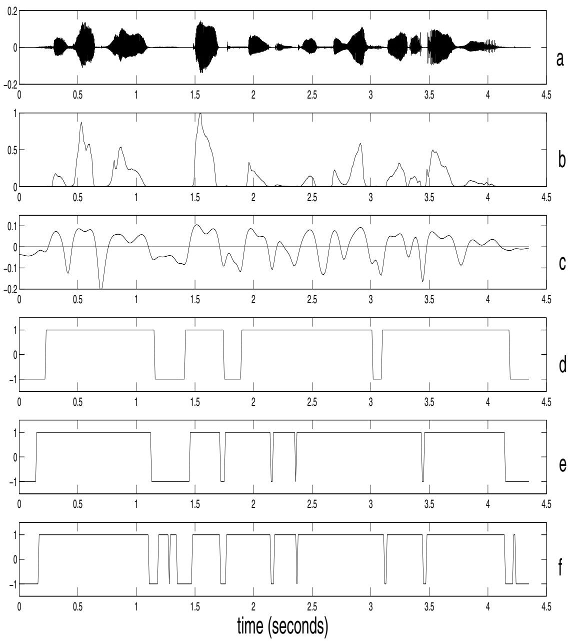

Fig. 1. Figures (a)-(d) describe the steps involved in performing VAD using group delay on a clean speech signal using WSF=3 ,$\gamma = 0.01$. (a) Clean speech signal (b) STE (c) Group delay with STE as magnitude spectrum (d) GD-VAD (e) Manual VAD (f) G.729B-VAD

2. Compute the zeroth lag $R_{M,GD}(0)$, of the cross-correlation between $V_{GD}(n)$ and $V_M(n)$ and normalize it with the length of the sequence.
3. Compute the zeroth lag $R_{M,G729}(0)$, of the cross-correlation between $V_{G729}(n)$ and $V_M(n)$ and normalize it with the length of the sequence.

C. Results

The normalized correlation of GD-VAD and G.729B VAD with the manual VAD for all the 432 speech files at a given SNR is computed as described in the previous subsection. It was observed that the average of the normalized correlation is significantly higher for the proposed method compared to G.729B at low SNR values. Figures 2, 3, 4 show plots of the normalized correlation values of GD-VAD and G.729B VAD for all the 432 speech signals corrupted with babble, white and pink noise respectively. From these figures it can be seen that the average normalized correlation of GD-VAD is much higher than that of G.729B VAD. This indicates a better frame-by-frame match with the manual VAD.

Also, Figure 5 shows a noisy speech signal at -15 dB corrupted by white noise with the corresponding manual VAD, G.729B VAD and GD-VAD. From this figure it can be seen that even at a very low SNR values, GD-VAD identifies speech and non-speech much better. The results given in Tables I, II, III further illustrate this.

It can be observed that the performance of G.729B VAD degrades as the SNR decreases. However, the performance of GD-VAD remains almost constant. This can be attributed to the fact that the expected energy is higher in high SNR regions for any noise level. Therefore, the STE of the signal, when viewed as a magnitude spectrum, indicates the presence of poles in these regions. Since the equivalent minimum phase group delay function resolves these poles much better than the magnitude spectrum [12], the high SNR regions in a signal are always resolved well at any noise level, indicating the robustness of group delay for VAD.

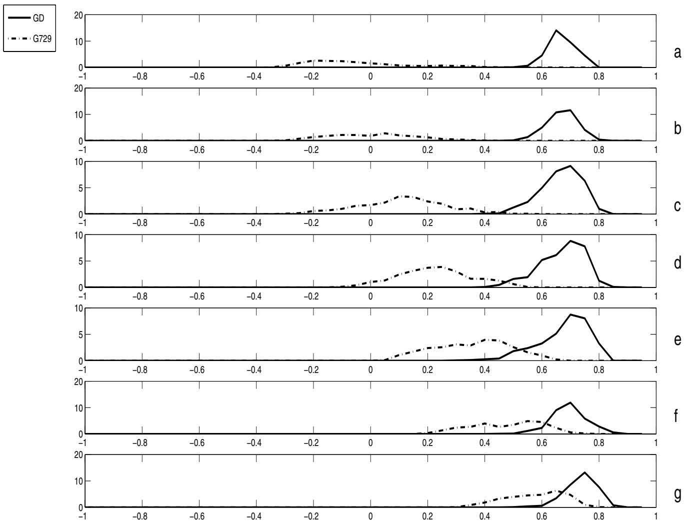

Fig. 2. GD-VAD and G.729B VAD correlation plots for 432 speech signals corrupted with **babble noise** at various SNR levels (dB): (a) -15, (b) -10, (c) -5, (d) 0, (e) 5, (f) 10, (g) 15

SNR(dB)	WSF	γ	$R_{M,GD}(0)$	$R_{M,G729}(0)$
-15	9	0.1	0.6675	-0.0769
-10	8	0.1	0.6677	-0.0123
-5	9	0.1	0.6627	0.0911
0	7	0.1	0.6647	0.2132
5	8	0.1	0.6703	0.3435
10	4	0.14	0.6907	0.4676
15	3	0.1	0.7340	0.5689

TABLE I

Normalized correlation values of GD-VAD and G.729B VAD for 432 speech signals corrupted with **babble noise** at various SNR levels

D. Effect of WSF and γ

Experiments were conducted for different values of the parameters, WSF and γ. It was found that γ does not make a significant difference in performance. However, as can be observed from Tables I, II, III, higher values of WSF yield better performance for noisier speech signals. This is due to WSF being inversely related to the cut-off frequency of the low-pass filter and hence eliminates the

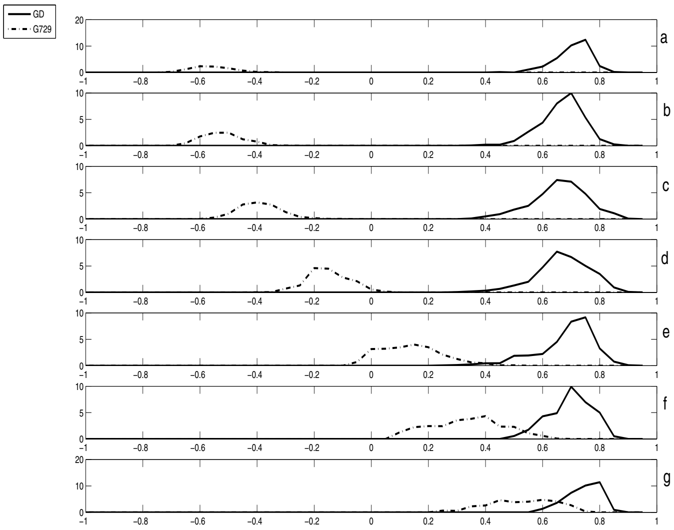

Fig. 3. GD-VAD and G.729B VAD correlation plots for 432 speech signals corrupted with **white noise** at various SNR levels (dB): (a) -15, (b) -10, (c) -5, (d) 0, (e) 5, (f) 10, (g) 15

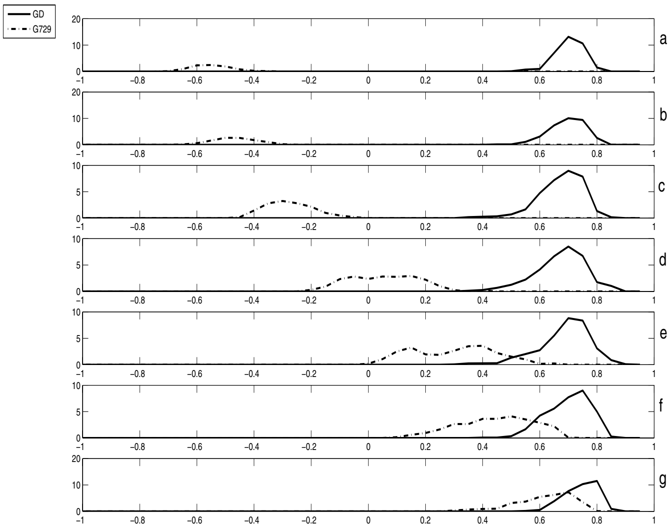

Fig. 4. GD-VAD and G.729B VAD correlation plots for 432 speech signals corrupted with **pink noise** at various SNR levels (dB): (a) -15, (b) -10, (c) -5, (d) 0, (e) 5, (f) 10, (g) 15

SNR(dB)	WSF	γ	$R_{M,GD}(0)$	$R_{M,G729}(0)$
-15	9	0.1	0.7036	-0.5697
-10	8	0.1	0.6628	-0.5326
-5	9	0.1	0.6516	-0.4001
0	9	0.1	0.6635	-0.1578
5	8	0.1	0.6750	0.1289
10	3	0.1	0.6931	0.3238
15	3	0.1	0.7404	0.5178

TABLE II
Normalized correlation values of GD-VAD and G.729B VAD for 432 speech signals corrupted with **white noise** at various SNR levels

SNR(dB)	WSF	γ	$R_{M,GD}(0)$	$R_{M,G729}(0)$
-15	9	0.2	0.7017	-0.5595
-10	8	0.1	0.6925	-0.4781
-5	8	0.1	0.6692	-0.2859
0	8	0.1	0.6663	0.0392
5	8	0.1	0.6797	0.2816
10	3	0.1	0.6946	0.4314
15	3	0.1	0.7424	0.6082

TABLE III
Normalized correlation values of GD-VAD and G.729B VAD for 432 speech signals corrupted with **pink noise** at various SNR levels

noise which contributes to the high frequency variations in the STE.

V. Conclusion

In this paper, a novel method for VAD in very low SNR is proposed. This method employs the group delay function to process the short-term energy of the speech signal. Experiments were performed with six different types of noise in a range of SNR values from -15 dB to 15 dB and the proposed method was compared with ITU's G.729 Annex B VAD algorithm. Zeroth lag of the normalized cross-correlation with manually marked VAD was used as a measure of performance of the algorithms. The proposed method performs significantly better than G.729 Annex B VAD algorithm for all the different types of noise at low SNR.

On the other hand, the proposed method needs to address the following issue: The G.729 Annex B algorithm makes VAD decisions frame by frame, whereas the proposed method makes VAD decisions per frame after processing the entire signal. Future work will investigate overcoming this. This however, is not a serious problem for many applications of VAD which process multiple frames (for eg., speech activity detection in ASR).

References

[1] Adil Benyassine, Eyal Shlomot, and Huan-Yu Su, "ITU Recommendation G.729 Annex B: A Silence Compression Scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Comm. Mag.*, pp. 64–73, 1997.

[2] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, pp. 1–3, 1999.

[3] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Process. Lett*, vol. 8, pp. 276–278, 2001.

[4] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Trans. Speech Audio Processing*, vol. 11, 2003.

[5] F. Beritelli, S. Casale, and A. Cavallaro, "A robust voice activity detector for wireless communications using soft computing," *IEEE Journal on Selected. Areas of Communication*, vol. 16, pp. 18181829, 1998.

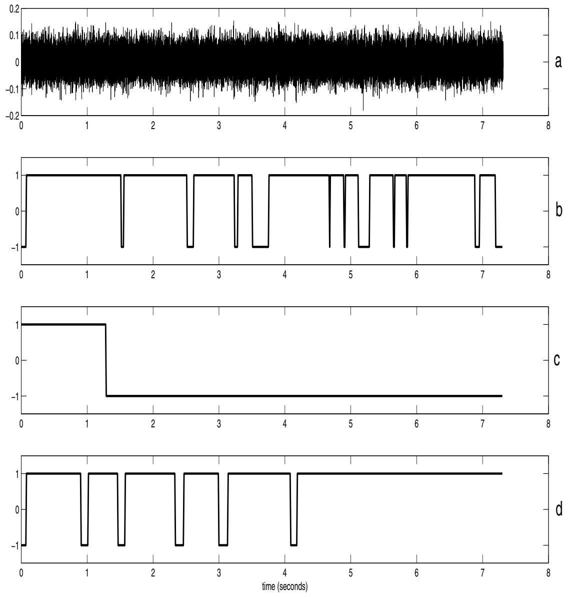

Fig. 5. (a) Noisy speech signal at -15dB SNR (b) Manual VAD (c) G.729B VAD (d) GD-VAD ($WSF = 9, \gamma = 0.1$)

[6] S. G. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 478482, 2000.

[7] H. Ozer and S. G. Tanyer, "A geometric algorithm for voice activity detection in nonstationary Gaussian noise," *EUSIPCO*, 1998.

[8] R. Tucker, "Voice activity detection using a periodicity measure," *Proc. IEEE*, vol. 139, pp. 377–380, 1992.

[9] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. Speech Audio Process*, vol. 9, pp. 217–231, 2001.

[10] C. Nikias and J. Mendel, "Signal processing with higher-order statistics," *IEEE Trans. Signal Processing*, vol. 41, pp. 10–38, 1993.

[11] Ke Li, M. N. S. Swamy, and M. Omair Ahmad, "An improved voice activity detection using higher order statistics," *IEEE Trans. Speech Audio Process*, vol. 13, 2005.

[12] V.Kamakshi Prasad, T.Nagarajan, and Hema A Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions," *Speech Comm*, vol. 42, pp. 429–446, 2004.

[13] L.R.Rabiner and R.W.Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Inc., Eaglewood Cliffs, New Jersey, 1978.

[14] T Nagarajan and Hema A Murthy, "Group Delay based segmentation of spontaneous speech into syllable-like units," *EURASIP Journal of Applied Signal Processing*, pp. 2614–2625, 2004.

[15] V. Kamakshi Prasad T. Nagarajan and Hema A. Murthy, "Minimum phase signal derived from root cepstrum," *IEE Electronics Letters*, vol. 39, 2003.

[16] Hema A Murthy and B.Yegnanarayana, "Formant extraction from minimum phase group delay function," *Speech Comm*, pp. 209–221, 1991.

[17] NTIS, *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, 1993.