Voice Activity Detection using Group Delay Processing on Buffered Short-term Energy

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

Abstract

In this paper, we present an algorithm for Voice Activity Detection (VAD) in speech signals using the minimum phase group delay function. The proposed method considers a buffer consisting of contiguous frames of the given signal and computes the short-term energy (STE) for that buffer. By appending a surrogate signal to STE and viewing the resultant signal as a positive part of the magnitude spectrum of an arbitrary signal, the minimum phase group delay function is computed. The group delay is then noise compensated and median filtered. The regions having positive group delay values are classified as speech and those with negative values are classified as noise. Experimental comparisons with the G.729 Annexe B VAD algorithm demonstrates significantly better performance for the proposed method, revealing that the algorithm is robust to noise.

1. Introduction

A Voice Activity Detection (VAD) algorithm distinguishes between speech and non-speech regions in a speech signal. It has several applications, including automatic speech recognition (ASR) and speech coding. VAD is a preprocessor stage to many such systems, and by selectively processing the output from VAD subsystem, the performance of the system is significantly improved. For example, in ASR, VAD improves the recognition accuracy significantly.

Traditionally, features like short-term energy, pitch etc. have been used for VAD. These are generally threshold based methods and do not perform well in low SNR conditions. The International Telecommunication Union recommends the G.729-Annexe B algorithm for VAD [1] and it is used in several speech coding systems. 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 techniques for VAD [2], [3], [4] model speech and noise as independent random variables. Methods for VAD based on the non-stationarity of noise were explored in [5], [6], [7]. Recently, some VAD algorithms exploit the fact that speech has certain higher order statistics (HOS) properties that are distinct from those of Gaussian noise. These algorithms model noisy-speech as a mixture of Gaussian and non-Gaussian processes [8], [9] . However, when noise is not Gaussian, or when unvoiced speech behaves like Gaussian noise, the distinction between speech and noise is ambiguous. Methods improving this have been proposed in [10].

Computing the minimum phase group delay function of the short-term energy of a signal has been explored for detecting syllable boundaries in [11]. The algorithm was extended to identify speech and non-speech regions in noisy conditions in [12]. However, the algorithm had a latency equal to the length of the signal. In this paper, we improve the algorithm further by reducing the latency as follows: (a) buffering contiguous frames of speech (b) handling the artifacts of buffering.

This paper is organized as follows: Section 2 reviews the group delay function. In Section 3, the proposed algorithm, GD-VAD is described. The performance comparison of the proposed method with G.729B is done in Section 4. Finally, we conclude in Section 5.

2. 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 (2) and (4) it can be observed that multiplication in the spectral domain corresponds to ad-

dition in the group delay domain. Also, group delay derived from a minimum phase signal is called the *minimum phase group delay function.* The additive and high-resolution properties of group delay functions are well established [13], [14], [11].

The VAD algorithm [12] uses a minimum phase signal derived from the short term energy (STE) function as if it were a magnitude spectrum. The group delay function of this minimum phase signal is then computed. The peaks in the group delay correspond to speech regions while the valleys represent non-speech regions.

3. The GD-VAD algorithm

The algorithm processes a buffer of the given speech signal as follows: For bth buffer, the short-term energy for each frame in that buffer is computed as $\acute{e}_b(m)$. This is appended with the surrogate signal, β rect(n) to form the sequence $\tilde{e}_b(m)$, where rect$(n) = u(n) - u(n-L)$, $u(n)$ is the unit step function and β is a scale factor determined empirically. This sequence is viewed as the positive part of the magnitude spectrum of an arbitrary signal and is converted into its minimum phase equivalent. The group delay of this signal is obtained. A noise-compensated group delay, $\tau_b^n(k)$, is then obtained by subtracting the maximum value of the group delay of the first few frames. Next, a median filtering on $\tau_b^n(k)$ using the current and past elements is performed to yield $\overline{\tau_b^n}(k)$. VAD-decision is done on $\overline{\tau_b^n}(k)$ as follows: positive values of $\overline{\tau_b^n}(k)$ are classified as speech regions and negative values are classified as non-speech. This algorithm is formally outlined below.

1. Given a speech signal $x(n)$, let us consider a buffer of contiguous frames. If the length of the buffer is B, then the number of buffers P = length of $x(n)/B$.

2. For each buffer b ($0 \leq b \leq P-1$), repeat steps 3 - 12:

3. Compute the STE, $\acute{e}_b(m)$, where $0 \leq m \leq B-1$.

4. **Append STE with β rect(n).** Form the sequence, $\tilde{e}_b(m)$

$$\tilde{e}_b(m) = \acute{e}_b(m) \qquad 0 \leq m \leq B-1 \quad (5)$$
$$\tilde{e}_b(m) = \beta\,\text{rect}(m) \quad B \leq m \leq B+L-1 \quad (6)$$
$$\tilde{e}_b(m) = 0 \qquad B+L-1 \leq m \leq M-1 \quad (7)$$

where rect$(n) = u(n) - u(n-L)$ and $M = 2^{\lceil \log_2(B+L) \rceil}$

5. Form the symmetric sequence, $\tilde{e}_{sb}(m)$

$$\tilde{e}_{sb}(m) = \tilde{e}_b(2M-m-1) \qquad M \leq m \leq 2M-1 \quad (8)$$

where $2M$ is the DFT order.

6. **Improving resolution using γ.** To improve the resolution, perform the following:

$$\breve{e}_{sb}(m) = \tilde{e}_{sb}(m)^{\gamma} \qquad 0 \leq m \leq 2M-1, \qquad 0 < \gamma \quad (9)$$

7. $\breve{e}_{sb}(m)$ is considered as a magnitude spectrum of an arbitrary signal of $2M$ points in $(-\pi, \pi)$ and is denoted by $E_b(k)$.

8. **Minimum phase equivalent.** Compute the IDFT of the function $E_b(k)$. The causal portion of the resulting sequence denoted by $e_b(l)$ is a minimum phase signal [15].

9. **Group delay computation.** Compute the group delay function [11], [14] of $e_b(l)w(l)$, where $w(l)$ is a cepstral lifter of length W_l as follows:

   - Compute the phase spectrum $\phi_b(k)$ of $e_b(l)w(l)$.
   - Compute the forward difference

   $$\tau_b(k) = \phi_b(k) - \phi_b(k-1) \qquad 1 \leq k \leq 2M-1$$

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

10. Compute noise-compensated group delay $\tau_b^n(k)$ as:

$$\tau_b^n(k) = \tau_b(k) - \tau_{\max} \qquad 0 \leq k \leq 2M-1 \quad (10)$$

where $\tau_{\max} = \max \tau_b(n)$ for $0 \leq n < T$ and T is an empirically determined threshold-index.

11. Perform median filtering on the noise-compensated group delay ($\tau_b^n(k)$) as:

$$\overline{\tau_b^n}(k) = median(\tau_b^n(k)) \qquad 0 \leq k \leq 2M-1 \quad (11)$$

where median(.) computes a 5-point median.

12. **VAD-decision.** If $\overline{\tau_b^n}(k) \geq 0$, classify frame as speech else if $\overline{\tau_b^n}(k) < 0$ classify frame as non-speech.

Window Scale Factor (WSF) is defined as $\frac{2M}{W_l}$. WSF and γ are used to control the resolution of the group delay. Figure 1(a) illustrates a noisy speech signal at 10 db SNR. Although VAD-decision for each frame is made buffer-wise, for illustration purposes, 1(b) shows the median-filtered noise-compensated group delay for all buffers. 1(c),(d) and (e) show the VAD-decisions by GD-VAD, G.729B VAD and manual VAD respectively.

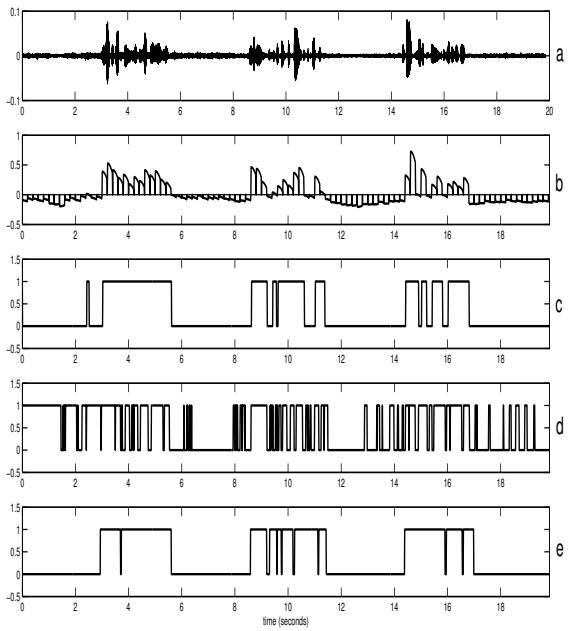

Figure 1: (a) Speech signal at 10dB SNR (b) Median-filtered noise-compensated group delay for the entire signal (c) GD-VAD (d) G.729B VAD (e) Manual VAD

4. Results and discussion

4.1. Experimental setup

432 speech files (216 female, 216 male) were obtained by concatenating sets of three individual speech utterances taken from TIMIT [16]. Additionally, to model the typical speech activity over a telephone conversation, silence was inserted so that the ratio of silence to speech is 60:40 [17]. These files were then low-pass filtered and resampled to 8kHz to conform to G.729B. Different types of noise (babble, pink and white) from the NOISEX-92 database were added resulting in fifteen test-sets, each having 0, 5, 10, 15 or 20 dB SNR respectively.

4.2. Results

Performance measures used for comparison between the proposed method and G.729B are hit ratio of speech (P_s) and hit ratio of noise (P_n) in comparison with manually marked VAD-decisions [17]. Further, front-end clipping and over-hang are ignored because hang-over schemes are not implemented. The performance metrics P_n and P_s are defined as:

$$P_n = \frac{\text{No. of non-speech frames from algorithm} \times 100}{\text{No. of non-speech frames in manual VAD}}$$

$$P_s = \frac{\text{No. of speech frames from algorithm} \times 100}{\text{No. of speech frames in manual VAD}}$$

In the course of the experiments, the following parameters were varied: signal SNR (0 to 20dB), buffer length B (step 2 in the algorithm), WSF (step 9) and γ (step 6). A constant $\gamma = 0.5$ and a buffer length of $B = 20$ frames (frame shift $=$ 10ms) yielded the best results. The results are tabulated below.

From the tables, it can be observed that the performance of GD-VAD is significantly better than G.729-VAD. As an example, figure (1) demonstrates the result. Also, there is an inverse relationship between SNR and WSF. This can be attributed to the fact that, at high noise conditions (low SNR), the short-term energy function fluctuated rapidly and consequently a higher WSF is mandated.

SNR	WSF	P_n^{GD}	P_s^{GD}	P_n^{G729B}	P_s^{G729B}
0	20	83.47	67.70	65.33	57.36
5	20	84.16	82.82	65.45	68.28
10	20	90.42	89.93	65.33	77.47
15	14	92.20	91.92	65.87	86.43
20	14	94.84	91.22	75.48	92.04

Table 1: Comparison of GD-VAD and G.729B-VAD for speech corrupted with **babble** noise.

SNR	WSF	P_n^{GD}	P_s^{GD}	P_n^{G729B}	P_s^{G729B}
0	24	95.95	75.05	89.50	60.52
5	22	95.60	85.69	88.84	71.48
10	16	95.76	90.79	89.38	81.43
15	14	95.59	92.56	88.40	88.65
20	14	96.58	91.26	93.22	94.49

Table 2: Comparison of GD-VAD and G.729B-VAD for speech corrupted with **pink** noise.

SNR	WSF	P_n^{GD}	P_s^{GD}	P_n^{G729B}	P_s^{G729B}
0	22	94.49	75.72	89.54	54.27
5	20	92.99	88.82	89.47	66.82
10	16	94.76	92.64	89.33	77.47
15	14	94.01	93.29	93.11	87.43
20	14	96.54	91.25	94.52	91.28

Table 3: Comparison of GD-VAD and G.729B-VAD for speech corrupted with **white** noise.

5. Conclusion

In this paper, we propose a method for VAD by group delay processing of the short-term energy of a given signal. Experiments were conducted on test data sets prepared from TIMIT using various types of noise at different SNR levels. The results are then compared with the G.729B VAD algorithm and is shown to perform significantly better. On the other hand, since the proposed

method is a buffered algorithm, it has a higher latency (200 ms) than G.729B (10 ms). This, however, is not an issue for applications like speech activity detection in automatic speech recognition.

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

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

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

[8] 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.

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

[10] 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.

[11] 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.

[12] Sree Hari Krishnan.P, R.Padmanabhan, and Hema A Murthy, "Robust voice activity detection using group delay functions," *IEEE ICIT*, 2006.

[13] B.Yegnanarayana, D.K.Saikia, and T.R.Krishnan, "Significance of group delay functions in signal reconstruction from spectral magnitude or phase," *IEEE Transactions on Acoustics, Speech and Signal Processing*, pp. 610–622, June 1984.

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

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

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

[17] Casale.S. Beritelli.F. and Ruggeri.G, "Performance evaluation and comparison of itu-t/etsi voice activity detectors," *Proceedings of ICASSP*, vol. 3, pp. 1425–1428, 2001.