A New Adaptive Long-Term Spectral Estimation Voice Activity Detector

Javier Ramírez, José C. Segura, Carmen Benítez,
Ángel de la Torre, Antonio Rubio

Department of Electronics and Computer Technology
University of Granada, Spain
javierrp@ugr.es

Abstract

This paper shows an efficient voice activity detector (VAD) that is based on the estimation of the long-term spectral divergence (LTSD) between noise and speech periods. The proposed method decomposes the input signal into overlapped speech frames, uses a sliding window to compute the long-term spectral envelope and measures the speech/non-speech LTSD, thus yielding a high discriminating decision rule and minimizing the average number of decision errors. In order to increase non-speech detection accuracy, the decision threshold is adapted to the measured noise energy while a controlled hang-over is activated only when the observed signal-to-noise ratio (SNR) is low. An exhaustive analysis of the proposed VAD is carried out using the AURORA TIdigits and SpeechDat-Car (SDC) databases. The proposed VAD is compared to the most commonly used ones in the field in terms of speech/non-speech detection and recognition performance. Experimental results demonstrate a sustained advantage over G.729, AMR and AFE VADs.

1. Introduction

The emerging applications of speech technologies (especially in mobile communications, robust speech recognition or digital hearing aids devices) often require a noise reduction scheme working in combination with a precise voice activity detector (VAD) [1]. There exist well known noise suppression algorithms [2], [3] such as Wiener filtering or spectral subtraction, that are widely used for robust speech recognition, and for which, the VAD is critical in attaining a high level of performance. These techniques estimate the noise spectrum during non-speech periods in order to compensate its harmful effect on the speech signal. For non-stationary noise environments, the VAD is even more critical since it is needed to update the constantly varying noise statistics. Thus, a correct classification of the incoming noisy speech signal is essential to track an accurate noise estimation and an efficient application of the noise suppression procedure. There exist also algorithms [4] that continually update the noise spectrum in order to prevent a misclassification of the speech signal causes a degradation of the enhanced signal. These techniques are faster in updating the noise but usually capture signal energy during speech periods, thus degrading the quality of the compensated speech signal. In this way, it is clearly better using an efficient VAD for most of the noise suppression systems and applications.

During the last decade different VAD methods have been proposed for several applications including mobile communication services [5], real-time speech transmission on the Internet [6] and noise reduction for digital hearing aids [7]. The detection principles are based fundamentally on the signal subband energy [8], its spectrum [9], [10], zero crossing rates (ZCR) [11], cepstral coefficients [12], Fuzzy rules [13], etc. This paper presents an efficient VAD with a high discriminating decision rule that is based on the estimation of the long-term spectral envelope. The VAD is compared to the most representative standards for voice activity detection such as the ITU G.729 [11] and ETSI AMR [14] and AFE [15]. Results were obtained for the AURORA databases and tasks [16], [17].

2. LTSE VAD Algorithm

The proposed speech/pause detection algorithm assumes that the most significant information for detecting voice activity on a noisy speech signal remains on the time-varying signal spectrum magnitude. Thus, the proposed VAD is based on the estimation of the Long-Term Spectral Envelope (LTSE). The decision rule is then formulated in terms of the Long-Term Spectral Divergence (LTSD) between speech and noise periods.

The algorithm can be described as follows. During a short initialization period, the mean noise spectrum $N(k)$ (*k*= 0, 1, ..., *NFFT*-1) is estimated by averaging the noise spectrum magnitude. After the initialization period, the LTSE VAD algorithm decomposes the input utterance into overlapped frames being their spectrum, namely $X(k,n)$, processed by means of a $(2N+1)$-frame window. The LTSD is obtained by computing the LTSE as:

$$LTSE(k) = max\left\{X(k,n+l)\right\}_{l=-N}^{l=+N} \quad (1)$$

where n is the frame for which the VAD decision is made and *k*= 0, 1, ..., *NFFT*-1, is the spectral band. The VAD decision rule is based on the LTSD that is calculated as the deviation of the LTSE respect to the noise spectrum defined to be:

$$LTSD = 10\log_{10}\left(\frac{1}{NFFT}\sum_{k=0}^{NFFT-1}\frac{LTSE^2(k)}{N^2(k)}\right) \quad (2)$$

Note that in order to reduce the computational overhead, the LTSD can be efficiently computed by taking advantage of the periodic nature of the LTSE and the noise.

The LTSD defined by equation (2) is a biased magnitude and needs to be compensated by a given *offset*. This value depends on the noise spectral variance and can be estimated during the initialization period or assumed to take a fixed value. The VAD makes the speech/non-speech decision comparing the unbiased LTSD to an adaptive threshold γ. The detection threshold is fixed during the VAD initialization according to the observed noise energy E. Optimal thresholds γ_0 and γ_1 for clean

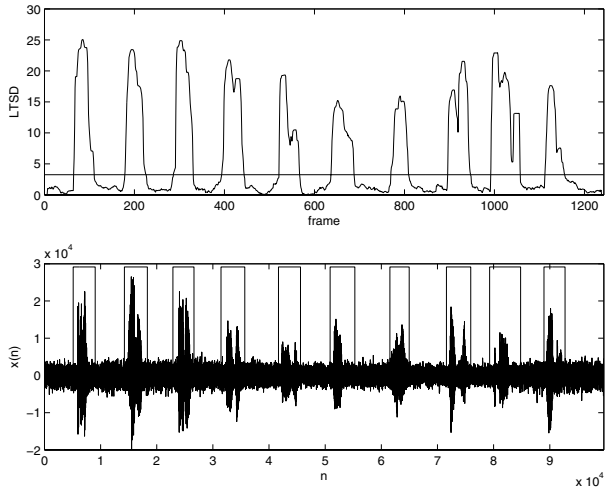

Figure 1: *VAD output for an utterance of the Spanish SpeechDat-Car database (recording conditions: high speed, good road, distant microphone).*

and high noise conditions, respectively, are defined and a linear VAD calibration curve is used. Thus, the optimum threshold γ is calculated as a function of the noise energy by:

$$\gamma = \begin{cases} \gamma_0 & E \leq E_0 \\ \frac{\gamma_0 - \gamma_1}{E_0 - E_1} E + \gamma_0 - \frac{\gamma_0 - \gamma_1}{1 - \frac{E_1}{E_0}} & E_0 < E < E_1 \\ \gamma_1 & E \geq E_1 \end{cases} \quad (3)$$

where E_0 and E_1 are the average noise energy for clean and high noise conditions, respectively. A high speech/non-speech discrimination is ensured with this model since speech pause detection is improved at high and medium SNR levels while maintaining a high precision detecting speech periods under high noise conditions.

The VAD is defined to be adaptive to time-varying noise environments with the following algorithm for updating the noise spectrum during non-speech periods being used:

$$N(k) = \alpha N(k) + (1 - \alpha) N_K(k) \quad (4)$$

where N_K is the average spectrum magnitude over a *K*-frame neighbourhood:

$$N_K(k) = \frac{1}{2K+1} \sum_{l=-K}^{K} X(k, n-l) \quad (5)$$

and *k*= 0, 1, ..., *NFFT/2*.

Finally, a hangover was found to be beneficial to maintain a high accuracy detecting speech periods at low SNR levels. If the LTSD achieves a given threshold $LTSD_0$, the hangover mechanism is turned off to increase speech pause detection when the noise level is low. Thus, the LTSE VAD yields an excellent classification of speech and pause periods. An example of the operation of the LTSE VAD on an utterance of the Spanish SpeechDat-Car database is shown in Fig. 1.

3. Experimental Framework

Several experiments were conducted using the AURORA databases and tasks [16]. This section evaluates the speech/non-speech discrimination as a function of the SNR level, provides

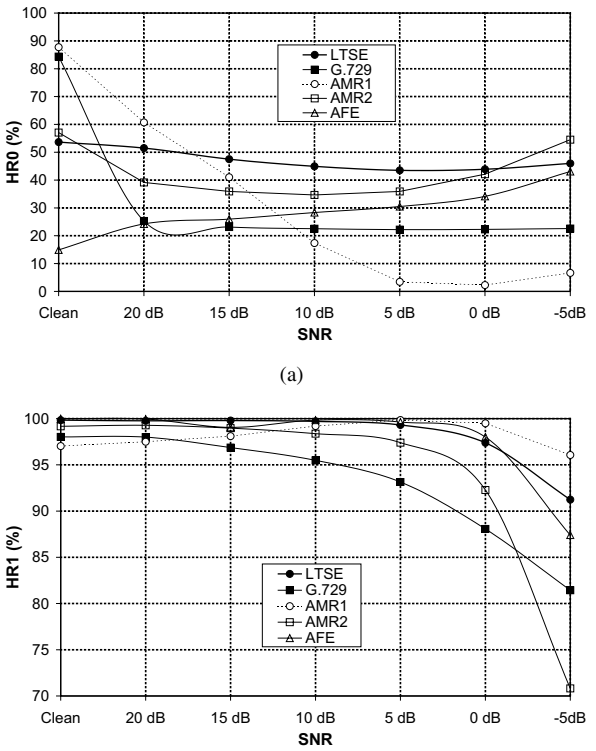

Figure 2: (a) *Non-speech hit-rate (HR0)*. (b) *Speech hit rate (HR1)*.

ROC curves for speech databases recorded under real conditions and compares speech recognition performance.

3.1. Speech/non-Speech Discrimination Analysis

First, the proposed VAD was evaluated in terms of the ability to discriminate between speech and pause periods. The clean TIdigits database was used to label each utterance as speech or pause frames for reference. VAD performance as a function of the SNR was assessed in terms of the pause and speech hit-rates (i.e., the fraction of all actual pause or speech frames that are correctly detected as pause or speech frames, respectively). Fig. 2 compares the proposed LTSE VAD to G.729, AMR and AFE VADs in terms of speech pause hit-rate (HR0, Fig. 2.a) and speech hit-rate (HR1, Fig. 2.b), both averaged for the entire set of noises, for clean conditions and for SNR levels ranging from 20 to -5 dB. The parameters used for the LTSE VAD were: N= 12, γ_0= 6, E_0= 30, γ_1= 2.5, E_1= 50, *offset*= 5, α= 0.95, K= 3, $LTSD_0$= 25, HO= 8 (hang-over length). Table 1 compares the VADs in terms of the average hit-rates. Thus, LTSE obtains the best behavior in detecting speech pauses with a 47.28 % HR0

Table 1: *Average speech/non-speech hit rates for SNR levels ranging from "clean" conditions to -5dB.*

	G.729	AMR1	AMR2	AFE	LTSE
HR0(%)	31.77	31.31	42.77	28.74	47.28
HR1(%)	93.00	98.18	93.76	97.70	98.15

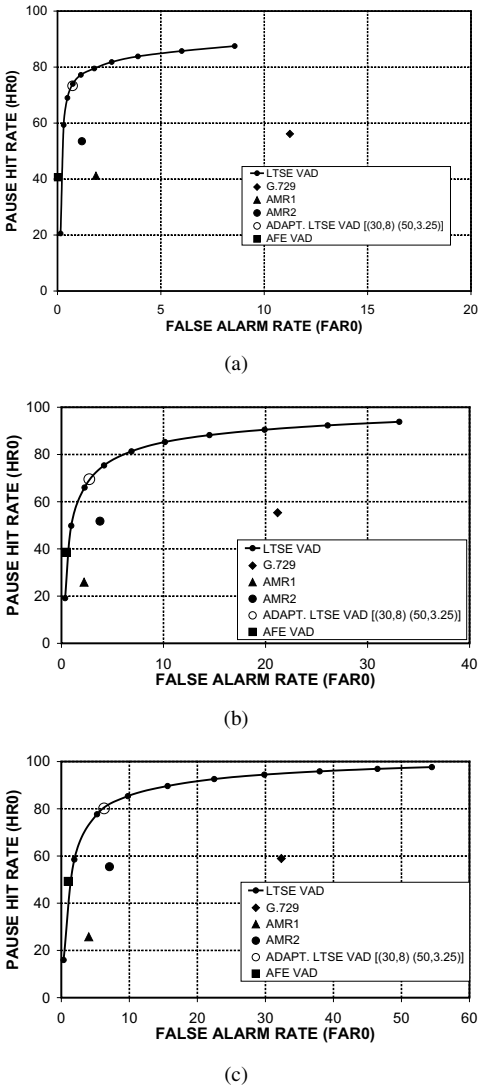

Figure 3: (a) *Stopped Car, Motor Running*. (b) *Town Traffic, Low Speed Rough Road*. (c) *High Speed, Good Road*.

average value, while the G.729, AMR1, AMR2 and AFE VADs yield 31.77 %, 31.31 %, 42.77 % and 28.74 %, respectively. On the other hand, the LTSE VAD is the most precise VAD in detecting speech periods exhibiting the slowest decay in performance at unfavorable noise conditions as shown in Fig. 2.b. Thus, LTSE attains a 98.15 % HR1 average value in speech detection while G.729, AMR1 and AMR2 VADs provide 93.00 %, 98.18 %, 93.76 % and 97.70 %, respectively. Although AMR1 seems to be well suited for speech detection at low SNRs, its extremely conservative behavior degrades its non-speech detection accuracy being HR0 less than 10 % below 10 dB, making it less useful in a practical speech processing system. Thus, considering together speech and pause hit-rates, the proposed VAD yielded the best results when compared to the most representative VADs tested.

3.2. VAD Comparison by means of the ROC Curves

An additional test was conducted to compare speech detection performance by means of the VAD ROC (Receiver Operating Characteristic) curve [18], a frequently used methodology based on the probabilities for false alarm and miss [8], that completely describes the VAD error rate. The Spanish SpeechDat-Car (SDC) database [19] was used in the analysis. This database contains 4914 recordings (files) from more than 160 speakers. Recordings from the close-talking microphone and from one of the distant microphones are included. As in the whole SDC database, the files are categorized into three noisy conditions (quiet, low noisy and highly noisy) depending on the driving conditions. Thus, recordings from the close-talking microphone are used in the analysis to label speech/pause frames for reference, while recordings from the distant microphone are used to evaluate the different VADs in terms of the ROC curves.

The speech pause hit rate (HR0) as a function of the false alarm rate (FAR0= 100-HR1) for $0 < \gamma \leq 10$ is shown in Fig. 3 for quiet, low and high noise conditions. The working point of the adaptive LTSE, G.729, AMR and the recently approved AFE VAD [15] are also included. It can be derived from these plots that the LTSE VAD yields the lowest false alarm rate for a fixed speech pause hit rate and also, the highest speech pause hit rate for a given false alarm rate. It is clearly shown that the ability of the adaptive LTSE VAD to tune the detection threshold enables working on the optimal point of the ROC curve for different noisy conditions. Thus, the adaptive LTSE VAD provides a sustained improvement in both speech pause hit rate and false alarm rate over G.729 and AMR VAD being the gains especially important over the G.729 VAD. LTSE VAD yields average absolute improvements in the speech pause hit rate of 17.54 %, 43.37 %, 20.76 % and 31.50 % over G.729, AMR1, AMR2 and AFE, respectively, while the average reduction in the false alarm rate is 27.78 %, -0.55 %, 0.76 % and -2.75 %. Thus, the proposed VAD yields the best speech pause detection accuracy, important reduction of the false alarm rate when compared to G.729, and comparable speech detection accuracy when compared to AMR VADs. Note that the 0.55 % increase in the false alarm rate over AMR1 is only motivated by the extremely conservative behavior of this VAD that only detects the 30.97 % of the real speech pauses while LTSE detects 74.34 % of the silence frames. It must be noted that the AFE VAD is only used in the standard [15] for frame-dropping and it has been planned to be conservative exhibiting poor speech pause detection accuracy, thus working on a low false alarm rate point of the ROC curve shown in Fig. 3.

3.3. Recognition Performance Analysis

Additionally, these improvements were corroborated when the VAD was integrated in a speech recognition system. The reference framework is the ETSI AURORA project for distributed speech recognition (DSR) [16], [17] and throughput is assessed in terms of the word accuracy (WAcc.). Table 2 shows the results obtained for two types of experiments conducted on the AURORA 2 (A2) and 3 (A3) databases: the effect of the VAD when (i) it is only used for Wiener filtering (WF), and (ii) it is applied for both, WF and removing non-speech periods (WF+frame-dropping(FD)). The best recognition performance is obtained when the proposed LTSE VAD is also used for FD. Thus, in clean training (CT) the relative improvements in the WAcc were 58.71 %, 49.36 % and 17.66 % over G.729, AMR1 and AMR2 VADs, respectively, while in multicondition training (MCT) the advantages were of up to

Table 2: Recognition performance results.

	Train/test		WAcc.(%)			
			LTSE	G.729	AMR1	AMR2
WF	A2	MCT	88.70	75.24	83.64	88.40
		CT	79.25	57.13	66.30	78.33
	A3	HM	68.62	67.31	64.94	65.44
		MM	81.70	77.66	76.34	76.58
		WM	93.18	92.21	92.58	92.71
WF+FD	A2	MCT	89.44	83.55	83.56	87.29
		CT	82.28	57.08	65.01	78.48
	A3	HM	83.69	63.35	76.47	79.53
		MM	86.16	67.66	81.29	82.37
		WM	95.10	88.81	94.93	94.91

35.81 %, 35.77 % and 16.92 %. Similar improvements were obtained for the experiments conducted on the Spanish, German and Finnish SDC databases for the three training/test modes defined (HM: high-mismatch, MM: medium-mismatch and WM: well matched). Again, the LTSE VAD provided the best results with 52.73 %, 44.27 % and 7.74 % average improvements for the different training/test modes and databases over G.729, AMR1 and AMR2, respectively, when the VAD is used for both WF and FD.

4. Conclusions

This paper showed an innovative LTSE-based VAD and analyzed its integration in a speech recognition system. The algorithm proposed is intended to mitigate the performance degradation suffered by most of the speech processing systems working under noise conditions. A complete analysis and comparison to existing standard VADs was carried out using the AURORA TIdigits and SDC databases. The proposed VAD yielded high levels of performance for different noises and SNR conditions, showed a clear advantage to the G.729, AMR1 and AMR2 VAD algorithms and it was preferred in applications where the speech signal is affected by undesired noise sources.

5. Acknowledgements

This work has been supported by the Spanish Government under TIC2001-3323 research project.

6. References

[1] Bouquin-Jeannes, R., L., Faucon, G., "Study of a Voice Activity Detector and its Influence on a Noise Reduction System", Speech Comm. (16), pp. 245–254, 1995.

[2] Boll, S. F., "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-27, pp. 113–120, Apr. 1979.

[3] Berouti, M., Schwartz, R., Makhoul, J., "Enhancement of Speech Corrupted by Acoustic Noise", Intl Conf. on Acoustics, Speech and Signal Processing, pp. 208–211, 1979.

[4] Martin, R., "An efficient algorithm to estimate the instantaneous SNR of speech signals", Eurospeech, vol. 1, 1993.

[5] Freeman, D., K., Cosier, G., Southcott, C. B., Boyd, I., "The Voice Activity Detector for the PAN-European Digital Cellular Mobile Telephone Service", Proc. of the International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 369–372, 1989.

[6] Sangwan, A., Chiranth, M. C., Jamadagni, H. S., Sah, R., Prasad, R. V., Gaurav, V., "VAD Techniques for Real-Time Speech Transmission on the Internet", 5th IEEE International Conference on High-Speed Networks and Multimedia Communications, pp. 46–50, 2002.

[7] Itoh, K., Mizushima, M.,"Environmental noise reduction based on speech/non-speech identification for hearing aids", International Conference on Acoustics, Speech, and Signal Processing, Vol. 1 , pp. 21–24, Apr. 1997.

[8] Marzinzik, M. and Kollmeier B., "Speech Pause Detection for Noise Spectrum Estimation by Tracking Power Envelope Dynamics", IEEE Trans. Speech and Audio Proc., 10(2):109–118, 2002.

[9] J. Sohn, N. S. Kim and W. Sung, "A statistical model-based voice activity detection", IEEE Signal Processing Letters, vol. 6, No. 1, pp. 1–3, 1999.

[10] Cho, Y. D., Al-Naimi, K., Kondoz, A., "A statistical model-based voice activity detection", Electronics Letters, vol. 37, No. 8, pp. 540–542, 2001.

[11] ITU-T Recommendation G.729 (Annex B): A Silence Compression Scheme for G.729, Optimized for Terminals Conforming to Recommendation V.70, ITU, 1996.

[12] Martin, A., Charlet, D., Mauuary, L., "Robust Speech/non-Speech Detection Using LDA Applied to MFCC", Intl. Conf. on Acoustics, Speech, and Signal Processing, Vol. 1 , pp. 237–240, 2001.

[13] Beritelli, F., Casale, S., Ruggeri, G., Serrano, S., "Performance Evaluation and Comparison of G.729/AMR/Fuzzy Voice Activity Detectors", IEEE Signal Processing Letters, Vol. 9, no. 3, pp. 85–88, 2002.

[14] ETSI EN 301 708 Recommendation: Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels, ETSI, Sophia Antipolis, Dec. 1999.

[15] ETSI ES 202 050 Recommendation: Speech processing, Transmission and Quality aspects (STQ);Distributed speech recognition;Advanced front-end feature extraction algorithm;Compression algorithms, 2000.

[16] Hirsch, H. G., and Pearce, D., "The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions", ISCA ITRW ASR2000: Automatic Speech Recognition: Challenges for the Next Millennium, Paris, France, Sept. 2000.

[17] ETSI ES 201 108 Recommendation: Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms, 2000.

[18] Madisetti, V., Williams, D. B., (Editors), Digital Signal Processing Handbook, CRC/IEEE Press, 1999.

[19] Moreno, A., *et al.*, "SpeechDat-Car: A Large Speech Database for Automotive Environments", Proc. II LREC, Jun. 2000.