VOICE ACTIVITY DETECTION WITH NOISE REDUCTION AND LONG-TERM SPECTRAL DIVERGENCE ESTIMATION

J. Ramírez, J. C. Segura, C. Benítez, A. de la Torre, A. Rubio

Dept. of Electronics and Computer Technology
University of Granada, Spain

ABSTRACT

This paper is mainly focussed on an improved voice activity detection algorithm employing long-term signal processing and maximum spectral component tracking. The benefits of this approach has been analyzed in a previous work with clear improvements in speech/non-speech discriminability and speech recognition performance in noisy environments. Two clear aspects are considered in this paper. The first one, which improves the performance of the VAD in low noise conditions, considers an adaptive length frame window to track the long-term spectral components. The second one reduces misclassification errors in high noisy environments by using a noise reduction stage before the long-term spectral tracking. Experimental results show clear improvements over different VAD methods in speech/pause discrimination and speech recognition performance. Particularly, the proposed VAD reported improvements in recognition rate when replaced the VADs of the ETSI Advanced Front-end (AFE) for distributed speech recognition (DSR).

1. INTRODUCTION

Modern applications of speech technology are demanding increased levels of performance in many areas. With the advent of wireless communications new speech services are becoming a reality with the development of modern robust speech processing technology. An important obstacle affecting most of the environments and applications is the environmental noise and its harmful effect on the system performance.

Most of the noise compensation algorithms often require to estimate the noise statistics by means of a precise voice activity detector (VAD). The detection task is not as trivial as it appears since the increasing level of background noise degrades the classifier effectiveness. During the last decade numerous researchers have studied different strategies for detecting speech in noise and the influence of the VAD decision on speech processing systems [1, 2]. Most of them have focussed on the development of robust algorithms, with special attention on the study and derivation of noise robust features and decision rules [3, 4, 5, 6].

This paper presents several improvements over a previous work on voice activity detection [7] that has been shown to be very effective for noise suppression and speech recognition in noisy environments. The algorithm assumes that the most significant information for detecting speech in noise remains on the time-varying signal spectrum. The main contributions of this paper are: *i*) the increased non-speech detection accuracy in low noise conditions by making the long-term window length adaptive to the measured noise energy, and *ii*) a noise reduction stage previous to tracking the long-term spectral envelope that improves the VAD effectiveness in high noise environments. The algorithm is evaluated on the context of the AURORA project and the recently approved Advanced Front-End (AFE) [8] for distributed speech recognition (DSR). Other standard VADs such as the ITU G.729 [9] or ETSI AMR [10] are also used as a reference.

This work was partly supported by the Spanish Government under the CICYT project TIC2001-3323.

2. LONG-TERM SPECTRAL ESTIMATION VAD

A block diagram of the proposed VAD is shown in Fig. 1. Noise reduction is performed first and the VAD decision is formulated on the de-noised signal. The noisy speech signal $x(n)$ is decomposed into 25-ms frames with a 10-ms window shift. Let $X(k,l)$ be the spectrum magnitude for the k-th band at frame l. The design of the noise reduction block is based on Wiener filter theory being its attenuation dependent on the signal-to-noise ratio (SNR) of the processed signal. Finally, the VAD decision is formulated in terms of the de-noised speech signal, being its spectrum $Y(k,l)$, processed by means of a $(2N+1)$-frame window.

2.1. Noise reduction

The noise reduction block consists of four stages: i) Spectrum smoothing. The power spectrum is averaged over two consecutive frames and two spectral bands. ii) Noise estimation. The noise spectrum $N_e(k)$ is updated by means of a 1^{st} order IIR filter based on the smoothed spectrum $Y_s(k,l)$, that is, $N_e(k) = \lambda N_e(k) + (1-\lambda)Y_s(k,l)$ where $\lambda = 0.99$. iii) Design of the Wiener filter (WF). First, the clean signal $S(k)$ is estimated by spectral subtraction:

$$S(k,l) = X_\beta S'(k,l) + (1 - X_\beta)\max(Y_s(k,l) - N_e(k), 0) \quad (1)$$

and the Wiener filter $H(k)$ is calculated by:

$$\begin{aligned} \eta(k,l) &= \max\left[\tfrac{S(k,l)}{N_e(k)}, \eta_{\min}\right] \\ H(k,l) &= \tfrac{\eta(k,l)}{1+\eta(k,l)} \end{aligned} \quad (2)$$

where $\eta_{\min}$ is selected so that the filter yield a 20 dB maximum attenuation, and X_β= 0.98. Finally, $S'(k,l)$, that is assumed to be zero at the beginning of the process, is defined to be:

$$S'(k,l) = \max[Y(k,l)H(k,l), 16] \quad (3)$$

The filter $H(k,l)$ is smoothed in order to eliminate rapid changes between neighbor frequencies that may often cause musical noise.

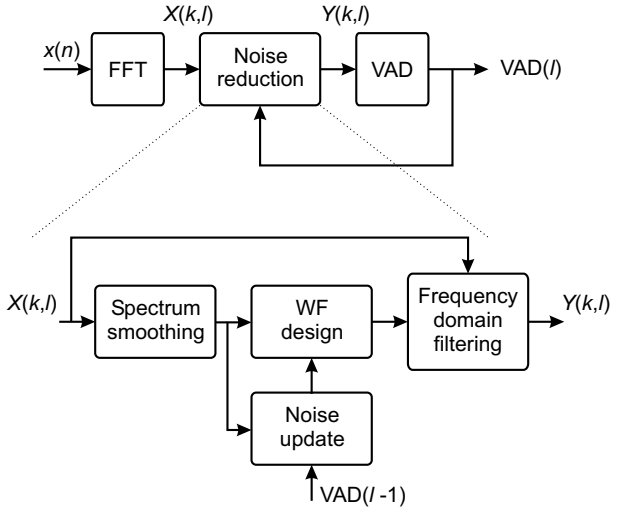


Fig. 1. *Block diagram of the VAD.*

The smoothing is performed by truncating the impulse response of the corresponding causal FIR filter to 17 taps using a Hanning window. iv) Frequency domain filtering. The filter is applied in the frequency domain to obtain the de-noised spectrum $Y(k,l)$.

2.2. VAD decision rule

Once the input speech has been de-noised, its spectrum magnitude $Y(k,l)$ is processed by means of a $(2N+1)$-frame window. Spectral changes around an N-frame neighborhood of the actual frame are analyzed using the N-order Long-Term Spectral Envelope (LTSE) that was defined in [7] as:

$$LTSE(k) = max\left\{Y(k,l+j)\right\}_{j=-N}^{j=+N} \tag{4}$$

where l is the actual frame for which the VAD decision is made and *k*= 0, 1, ..., *NFFT*-1, is the spectral band.

Note that the noise suppression block have to perform the noise reduction of the block $\{X(k,l-N)$, $X(k,l-N+1)$, ..., $X(k,l-1)$, $X(k,l)$, $X(k,l+1)$, ..., $X(k,l+N)\}$ before the LTSE at the l-th frame can be computed. This is carried out as follows. During the initialization, the noise suppression algorithm is applied to the first $2N+1$ frames and, in each iteration, the $(l+N+1)$-th frame is de-noised, so that $Y(k,l+N+1)$ become available for the next iteration.

The VAD decision rule is formulated in terms of the long-term spectral divergence (LTSD) calculated as the deviation of the LTSE respect to the residual noise spectrum $N(k)$ and defined by:

$$LTSD = 10\log_{10}\left(\frac{1}{NFFT}\sum_{k=0}^{NFFT-1}\frac{LTSE^2(k)}{N^2(k)}\right) \tag{5}$$

If the LTSD is greater than an adaptive threshold γ, the actual frame is classified as speech, otherwise it is classified as non-speech. A hangover delays the speech to non-speech transition in order to prevent low-energy word endings being misclassified as silences. On the other hand, if the LTSD achieves a given threshold $LTSD_0$, the hangover algorithm is turned off to improve non-speech detection accuracy in low noise environments.

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:

$$N(k) = \alpha N(k) + (1-\alpha)N_K(k) \tag{6}$$

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

$$N_K(k) = \frac{1}{2K+1}\sum_{j=-K}^{K} Y(k,n-j) \tag{7}$$

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

2.3. Initialization of the algorithm

For the initialization of the algorithm, the first frames of the input utterance are assumed to be non-speech and the decision threshold γ and N are adapted to the measured noise energy E by:

$$\begin{array}{l}\gamma = \frac{\gamma_0-\gamma_1}{E_0-E_1}E + \gamma_0 - \frac{\gamma_0-\gamma_1}{1-E_1/E_0} \\ N = \text{round}\left[\frac{N_0-N_1}{E_0-E_1}E + N_0 - \frac{N_0-N_1}{1-E_1/E_0}\right]\end{array} \tag{8}$$

where E_0 and E_1 are the average noise energy for clean and high noise conditions, respectively. Since γ and N are critical parameters for the algorithm, they are restricted to be bounded in the intervals $[\gamma_0,\gamma_1]$ and $[N_0,N_1]$, respectively.

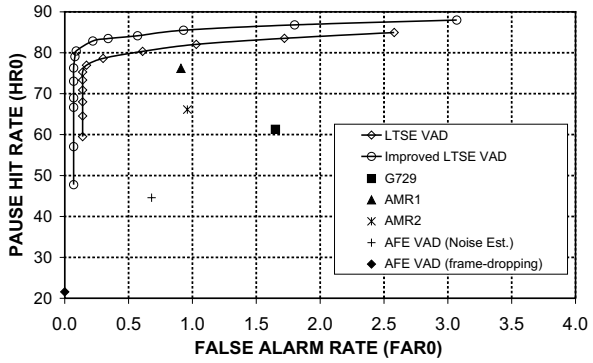
3. EXPERIMENTAL FRAMEWORK

Several experiments were conducted to evaluate the performance of the VAD. The analysis is focused on the assessment of misclassification errors at different SNR levels and the influence of the VAD decision on an automatic speech recognition (ASR) system.

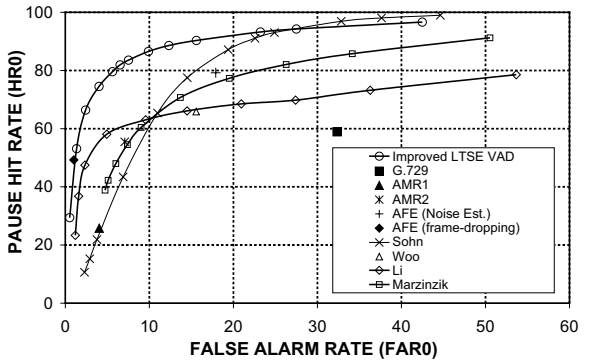
3.1. Receiver operating characteristics (ROC) curves

The ROC curves are frequently used to completely describe the VAD error rate. The AURORA subset of the original Spanish SpeechDat-Car (SDC) database [11] was used in this analysis. This database contains 4914 recordings using close-talking and distant microphones from more than 160 speakers. As in the whole SDC database, the files are categorized into three noisy conditions: quiet, low noisy and highly noisy conditions, which represent different driving conditions with average SNR values between 25dB, and 5dB. The non-speech hit rate (HR0) and the false alarm rate (FAR0= 100-HR1) were determined in each noise condition being the "real" speech frames and "real" speech pauses determined by hand-labelling the database on the close-talking microphone. The parameters used for the VAD were: N_0= 3 N_1= 6, α= 0.95, K= 3, $LTSD_0$= 40, HO= 3 (hang-over length).

Fig. 2.a shows the ROC curve for recordings from the close-talking microphone with low noise. The working points of the G.729, AMR and AFE VADs are also included. It can be derived from the figure that the improvements considered in this work yield higher speech/non-speech discrimination. This fact is mainly motivated by using a shorter frame window for low noise environments. Fig. 2. b shows the ROC curve for recordings from the distant microphone in high noisy conditions. It also shows improvements in detection accuracy over standard VADs such as G.729, AMR and AFE and over a representative set of recently reported VAD algorithms [3, 6, 4, 5].



(a)



(b)

Fig. 2. *ROC curves. (a) Close talking microphone (stopped car, motor running). (b) Distant microphone (high speed, good road).*

It can be concluded from these results that: i) The working point of the G.729 VAD shifts to the right in the ROC space with decreasing SNR. ii) AMR1 works on a low false alarm rate point of the ROC space but exhibits poor non-speech hit rate. iii) AMR2 yields clear advantages over G.729 and AMR1 exhibiting important reduction of the false alarm rate when compared to G.729 and increased non-speech hit rate over AMR1. iv) The VAD used in the AFE for noise estimation yields good non-speech detection accuracy but works on a high false alarm rate point on the ROC space. It suffers from rapid performance degradation when the driving conditions get noisier. On the other hand, the VAD used in the AFE for frame-dropping has been planned to be conservative since it is only used in the DSR standard for frame-dropping. Thus, it exhibits poor non-speech detection accuracy working on a low false alarm rate point of the ROC space. Among all the VAD examined, our VAD yields the lowest false alarm rate for a fixed non-speech hit rate and also, the highest non-speech hit rate for a given false alarm rate. The ability of the adaptive LTSE VAD to tune the detection threshold by means the algorithm described in Eq. 8 enables working on the optimal point of the ROC curve for different noisy conditions.

3.2. Speech recognition performance

Although the ROC curves are effective to evaluate a given algorithm, the influence of the VAD in an ASR system was also studied. The reference framework (Base) is the ETSI AURORA project for distributed speech recognition [12] while the recognizer is based on the HTK (Hidden Markov Model Toolkit) software package [13]. The task consists on recognizing connected digits which are modelled as whole word HMMs (Hidden Markov Models) with the following parameters: 16 states per word, simple left-to-right models, mixture of 3 Gaussians per state while speech pause models consist of 3 states with a mixture of 6 Gaussians per state. The 39-parameter feature vector consists of 12 cepstral coefficients (without the zero-order coefficient), the logarithmic frame energy plus the corresponding delta and acceleration coefficients. Two training modes are defined for the experiments conducted on the AURORA-2 database: *i*) training on clean data only (Clean Training), and *ii*) training on clean and noisy data (Multi-Condition Training).

The influence of the VAD decision on the performance of a feature extraction scheme incorporating Wiener filtering (WF) as noise suppression method and non-speech frame-dropping (FD) to the Base system [12] was assessed. Table 1 shows the recognition results as a function of the SNR for the Base system and for the different VADs that were incorporated to the feature extraction algorithm. These results are averaged over the three test sets of the AURORA-2 recognition experiments. An estimation of the 95% confidence interval (CI) is also provided. Notice that, particularly, for the recognition experiments based on the AFE VADs, we have used the same configuration used in the standard [8] which present different VADs for WF and FD. The proposed VAD outperforms the standard G.729, AMR1, AMR2 and AFE VADs in both clean and multi condition training/testing experiments.

Table 2 compares recently reported VAD algorithms to the proposed one in terms of the average word accuracy for clean and multicondition training/test experiments. The proposed algorithm also outperforms the VADs used as a reference being the one that is closer to the "ideal" hand-labelled speech recognition performance.

Finally, in order to compare the proposed method to the best available results, the VADs of the full AFE standard (including both the noise estimation and frame dropping VADs) were replaced by the proposed LTSE VAD and the AURORA recognition experiments were conducted. The results are shown in Table 3. The word error rate is reduced from 13.07% to 12.42% for the clean training experiments and from 8.14% to 7.88% in multicondition when the VADs of the original AFE are replaced by the proposed VAD.

4. CONCLUSIONS

This paper has shown an improved VAD algorithm for increasing speech detection robustness in noisy environments and the performance of speech recognition systems. The VAD is based on the estimation of the long-term spectral envelope and the measure of the spectral divergence between speech and noise. Two improvements have been considered over the base system. The first one, which improves the performance of the VAD in low noise conditions, considers a variable length frame window to track the long-term spectral components. The second one reduces misclassification errors in high noisy environments by using a noise reduction stage before the long-term spectral tracking. With this and other innovations the proposed VAD has demonstrated an enhanced ability to discriminate speech and silences and to be well suited for robust speech recognition.

Table 1. *Average Word Accuracy for the AURORA-2 database.*

	Base	**Multicondition**					**Clean condition**				
VAD used	None	G.729	AMR1	AMR2	AFE	**Proposed**	G.729	AMR1	AMR2	AFE	**Proposed**
Clean	99.03	97.50	96.67	98.12	98.39	98.83	98.41	97.87	98.63	98.78	99.11
20 dB	94.19	96.05	96.90	97.57	97.98	98.40	83.46	96.83	96.72	97.82	98.03
15 dB	85.41	94.82	95.52	96.58	96.94	97.59	71.76	92.03	93.76	95.28	96.44
10 dB	66.19	91.23	91.76	93.80	93.63	95.49	59.05	71.65	86.36	88.67	91.51
5 dB	39.28	81.14	80.24	85.72	85.32	88.49	43.52	40.66	70.97	71.55	77.25
0 dB	17.38	54.50	53.36	62.81	63.89	67.49	27.63	23.88	44.58	41.78	49.27
-5 dB	8.65	23.73	23.29	27.92	30.80	33.36	14.94	14.05	18.87	16.23	22.90
Average	60.49	83.55	83.56	87.29	87.55	**89.49**	57.08	65.01	78.48	79.02	**82.50**
C.I. (95%)	±0.24	±0.18	±0.18	±0.16	±0.16	±0.15	±0.24	±0.23	±0.20	±0.20	±0.19

Table 2. *Average Word Accuracy for the AURORA-2 database.*

VAD Used	**Clean**	**Multi**	**Average**
Woo	76.64	85.54	81.09
Li	78.63	85.58	82.11
Marzinzik	82.15	88.32	85.23
Sohn	79.49	88.11	83.80
Proposed	82.50	89.49	86.00
Hand-labelled	84.83	88.88	86.86

Table 3. *Average Word Accuracy for the AURORA-2 database.*

	Clean training		**Multicondition training**	
	AFE	AFE + LTSE	AFE	AFE + LTSE
Set A	87.51	88.12	92.29	92.55
Set B	87.06	87.79	92.10	92.41
Set C	85.42	86.10	90.51	90.70
Average	86.93	**87.58**	91.86	**92.12**

5. REFERENCES

[1] L. Karray and A. Martin, "Towards improving speech detection robustness for speech recognition in adverse environments," *Speech Communitation*, vol. 40, no. 3, pp. 261–276, 2003.

[2] A. Sangwan, M. C. Chiranth, H. S. Jamadagni, R. Sah, R. V. Prasad, and V. Gaurav, "VAD techniques for real-time speech transmission on the internet," in *IEEE International Conference on High-Speed Networks and Multimedia Communications*, 2002, pp. 46–50.

[3] K. Woo, T. Yang, K. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.

[4] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 341–351, 2002.

[5] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 16, no. 1, pp. 1–3, 1999.

[6] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 146–157, 2002.

[7] J. Ramírez, J. C. Segura, M. C. Benítez, A. de la Torre, and A. Rubio, "A new adaptive long-term spectral estimation voice activity detector," in *Proc. of EUROSPEECH 2003*, Geneva, Switzerland, September 2003, pp. 3041–3044.

[8] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," *ETSI ES 201 108 Recommendation*, 2002.

[9] ITU, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," *ITU-T Recommendation G.729-Annex B*, 1996.

[10] ETSI, "Voice activity detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels," *ETSI EN 301 708 Recommendation*, 1999.

[11] A. Moreno, L. Borge, D. Christoph, R. Gael, C. Khalid, E. Stephan, and A. Jeffrey, "SpeechDat-Car: A Large Speech Database for Automotive Environments," in *Proceedings of the II LREC Conference*, 2000.

[12] ETSI, "Speech processing, transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms," *ETSI ES 201 108 Recommendation*, 2000.

[13] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University, 1997.