A NEW VOICE ACTIVITY DETECTOR USING SUBBAND ORDER-STATISTICS FILTERS FOR ROBUST SPEECH RECOGNITION

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

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

ABSTRACT

Currently, there are technology barriers inhibiting speech processing systems working under extreme noisy conditions. The emerging applications of speech technology, especially in the fields of wireless communications, digital hearing aids or speech recognition, are some examples of such systems often requiring a noise reduction technique in combination with a precise voice activity detector (VAD). This paper presents a new VAD for improving speech detection robustness in noisy environments and the performance of speech recognition systems. The algorithm uses long-term information about the speech signal to formulate the decision rule and estimates the subband SNR using specialized order statistics filters (OSFs). The proposed algorithm is compared to the most commonly used VADs in the field, in terms of speech/non-speech discrimination and also in terms of recognition performance when the VAD is used in an automatic speech recognition (ASR) system. Experimental results demonstrate a sustained advantage over different VAD methods including standard VADs such as G.729 and AMR which are used as a reference, the VADs of the Advanced Front-End (AFE) for distributed speech recognition (DSR), and recently reported algorithms.

1. INTRODUCTION

Speech/non-speech detection is an unsolved problem affecting to numerous applications. The classification task is not as trivial as it appears and most of the voice activity detection (VAD) algorithms often fail when the level of background noise increases. 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. There exist well known noise suppression algorithms [1, 2], such as Wiener filtering (WF) 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. The VAD is more critical for non-stationary noise environments since it is needed to update the constantly varying noise statistics affecting a misclassification error strongly to the system performance. A representative set of recently published VAD methods formulates the decision rule on a frame by frame basis using instantaneous measures of the divergence distance between speech and noise [3, 4]. It has been shown recently that VAD robustness can be improved by using long-term spectral information to formulate the decision rule [5]. This paper explores a new strategy for detecting speech in noise using order statistics filters (OSFs) for the estimation of the subband speech/non-speech divergence.

2. SPEECH/NON-SPEECH DETECTION ALGORITHM

The algorithm is stated as follows. The input signal $x(n)$ is decomposed in 25-ms overlapped frames with a 10-ms window shift. The log-energies for the m-th frame, $E(m,k)$, in K subbands $(k = 0, 1, ..., K-1)$, are computed by means of:

$$\begin{array}{c} E(m,k) = \log\left(\frac{K}{L}\sum\limits_{l=l_k}^{l_{k+1}-1} X_m(l)\right) \\ l_k = \lfloor \frac{L}{2K} k \rfloor \;\; k = 0, 1, ..., K-1 \end{array} \tag{1}$$

where $X_m(l)$ (l=0, 1, …, L-1) is the power spectrum magnitude.

The algorithm uses two order statistics filters (OSFs) for the multi-band quantile (MBQ) SNR estimation. The implementation of both OSFs is based on a sequence of $2N$+1 log-energy values $\{E(m-N,k), \ldots, E(m,k), \ldots, E(m+N,k)\}$ around the frame to be analyzed. The r-th order statistics of this sequence, $E_{(r)}(m,k)$, is defined as the $r-$th largest number in algebraic order. For the estimation of the noise level in each subband, a median filter is used. A second OSF estimates the subband signal energy by means of:

$$Q_p(m,k) = (1-f)E_{(l)}(m,k) + fE_{(l+1)}(m,k) \tag{2}$$

where $Q_p(m,k)$ is the p sampling quantile, $l = \lfloor 2pN \rfloor$ and f=2pN-l. Finally, the SNR in each subband is measured by:

$$QSNR(m,k) = Q_p(m,k) - E_N(k) \tag{3}$$

where the sampling quantile p= 0.9 is selected as a good estimation of the subband spectral envelope. Finally, the decision rule is formulated in terms of the average subband SNR:

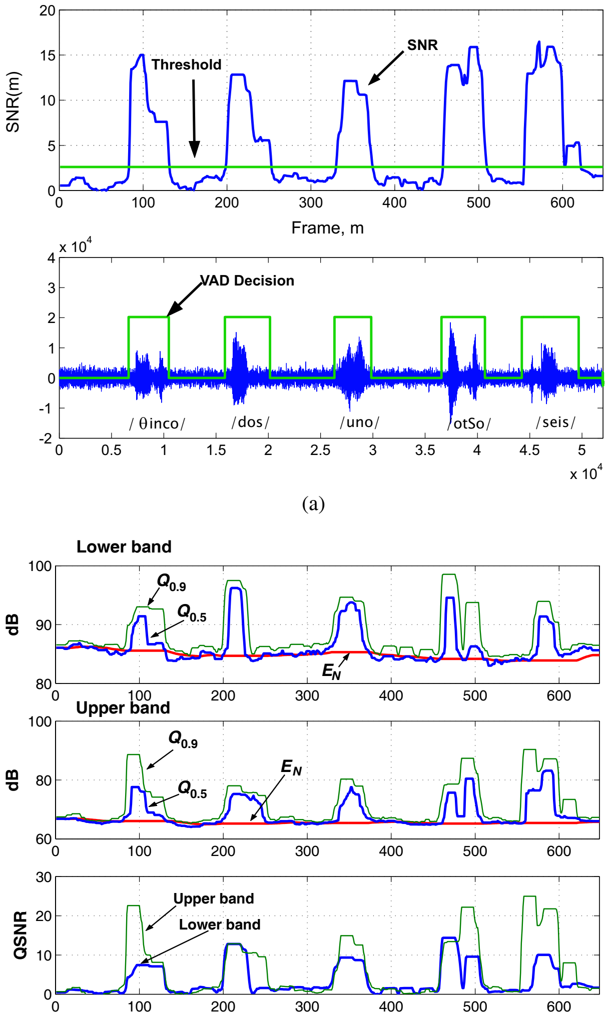
$$SNR(m) = \frac{1}{K}\sum_{k=0}^{K-1} QSNR(m,k) \tag{4}$$

For the initialization of the algorithm, the first N frames are assumed to be non-speech frames and the noise level in the k-th band, $E_N(k)$, is estimated as the median of the set $\{E(0,k), E(1,k), \ldots, E(N\text{-}1,k)\}$.

Thus, if the SNR is greater than the threshold η, the actual frame is classified as speech, otherwise it is classified as non-speech. The threshold is made adaptive to the measured full-band noise energy E in order to select the optimum working point for

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

0-7803-8484-9/04/$20.00 ©2004 IEEE



(a)

(b)

Fig. 1. *Operation of the VAD on an utterance of Spanish SDC database. (a) SNR and VAD Decision. (b) Subband SNR.*

different SNR conditions. The threshold is linearly decreased with the increasing noise level:

$$\eta = \min\left[\max\left[\frac{\eta_0 - \eta_1}{E_0 - E_1} E + \eta_0 - \frac{\eta_0 - \eta_1}{1 - E_1/E_0}, \eta_1\right], \eta_0\right] \quad (5)$$

between (E_0, η_0) and (E_1, η_1) for clean and high noisy conditions defined by E_0 and E_1, respectively. Finally, to track non-stationary noisy environments, the noise levels are updated during non-speech periods using a 1^{st} order IIR filter:

$$\begin{aligned} E_N(k) &= \alpha E_N(k) + (1-\alpha) Q_{0.5}(m,k) \\ k &= \ 0, \ 1, \ \ldots, \ K-1 \end{aligned} \quad (6)$$

being $Q_{0.5}(m,k)$ the output of the median filter and α= 0.97 was experimentally selected.

Fig. 1 shows the operation of the proposed VAD on an utterance of the Spanish SpeechDat-Car (SDC) database. In the example, K= 2 subbands are used being clearly shown how the SNR in the upper and lower band yields improved speech/non-speech discrimination of fricative sounds by giving complementary information.

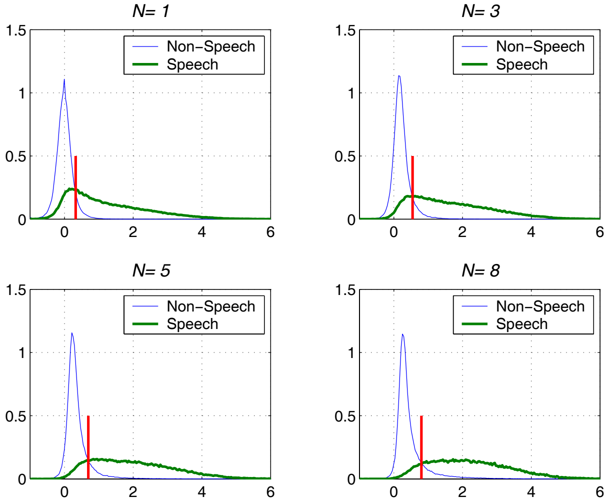


Fig. 2. *Effect of the window length on the Speech/non-Speech distributions.*

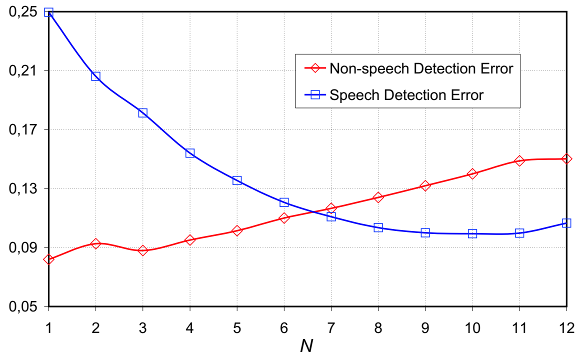


Fig. 3. *Speech and non-Speech detection error as a function of the window length.*

3. DISTRIBUTIONS OF SPEECH AND SILENCE

In order to clarify the motivations for the algorithm proposed, the distributions of the SNR defined by Eq. 4 as a function of the long-term window length (N) were studied. A hand-labelled version of the Spanish SDC database was used in the analysis. This database contains recordings from close-talking and distant microphones at different driving conditions: a) stopped car, motor running, b) town traffic, low speed, rough road and c) high speed, good road. The most unfavourable noise environment (i.e. high speed, good road) was selected and recordings from the distant microphone were considered. Thus, the N-order SNR was measured during speech and non-speech periods, and the histogram and probability distributions were built. Fig. 2 shows the distributions of speech and noise for N= 1, 3, 5 and 8. It is derived from Fig. 2 that speech and noise distributions are better separated when the order of the long-term window increases. This fact makes the VAD more robust against environmental noise since misclassification errors are reduced. The distribution of noise is highly confined around the mean value exhibiting reduced variance. Thus, the probability of detect noise as speech is reduced. On the other hand, the distribution of speech is shifted to the right as the window length increases being also reduced the probability of detecting speech as noise. This fact can be corroborated by calculating the classification error of speech and noise for an optimal Bayes classifier. Fig. 3 shows the classification errors as a function of the window length N. The speech classification error is approximately reduced by half from 25% to 10% when the order of the VAD is increased from 1 to 8 frames. This is motivated by the separation of the distributions that takes place when N is increased as shown in Fig. 2. On the other hand, the increased speech detection robustness is only prejudiced by a moderate increase of the non-speech speech detection error. According to Fig. 3, the optimal value of the order of the VAD would be N= 8. As a conclusion, a long-term measure of the SNR is beneficial for VAD since it significantly reduces misclassification errors.

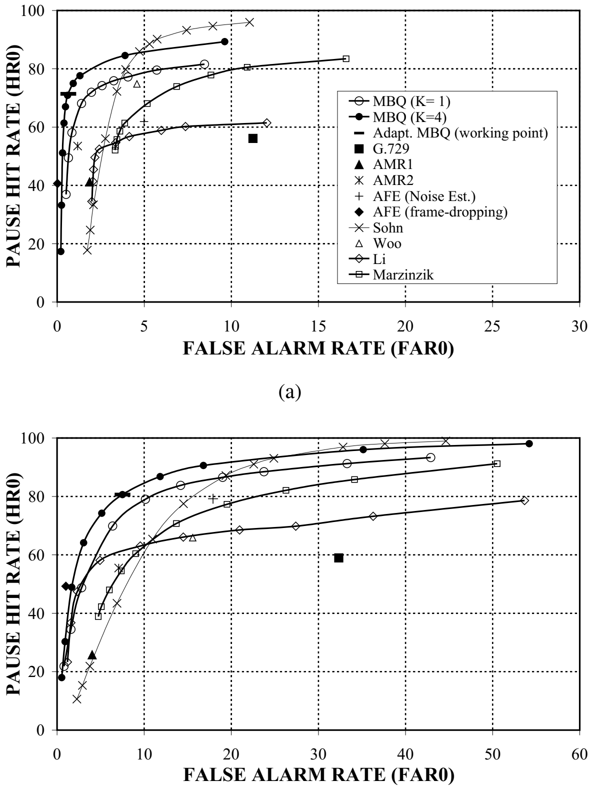
4. EXPERIMENTAL RESULTS

Several experiments were conducted for the evaluation of the proposed VAD. First, misclassification errors were studied at different SNR levels by means of the Receiver Operating Characteristics (ROC) curves. Second, the influence of the VAD decision on a speech recognition system was assessed. G.729 [6], AMR [7] and AFE [8] standards, as well as VAD algorithms recently reported by Woo [4], Li [9], Sohn [3] and Marzinzik [10] were used for reference.

4.1. Analysis of the ROC curves

The AURORA subset of the Spanish SDC database was used in this analysis. It was hand-labelled on the close talking microphone to obtain the speech/non-speech hit rates, HR1 and HR0, respectively. Fig. 4 shows the trade-off between speech pause hit rate and false alarm rate (FAR0= 100-HR1) for different driving (noisy) conditions, as the threshold varies. The optimum length of the block was N= 8 frames and the log-energies $E(m,k)$ were computed using L= 256 points for the FFT. The selection of the number of subbands is influenced by the trade-off between computational complexity and VAD performance. Using K= 4 subbands significantly increases the effectiveness of the proposed VAD. This fact is motivated by a shift up and to the left of the ROC curve when the number of subbands is increased. The adaptive MBQ VAD defined by thresholds η_0= 0.85 dB for E_0= 30 dB and η_1= 0.7 dB for E_1= 50 dB, according to Eq. 5, enables working on the optimal point when the SNR varies from 25 to 5 dB.

The proposed VAD works with lower false alarm rate and higher speech pause hit rate when compared to standards G.729 [6], AMR [7] and AFE [8](including the VADs used for noise estimation and frame-dropping) and the Sohn's [3], Woo's [4], Li's [9] and Marzinzik's [10] algorithms. The benefits are especially important over G.729 which is used along with a speech codec for discontinuous transmission, and the Li's algorithm that is based on an optimum linear filter for edge detection. The proposed VAD also improves Marzinzik's VAD that tracks the power spectral envelopes. There is a point where the Sohn's ROC curve starts being up MBQ VAD in the ROC space. However, this area is far from being optimum for most of the applications since HR1 is less than 95% and 75% for 12 and 5 dB SNRs, respectively, and excessive speech frames would be lost.



(a)

(b)

Fig. 4. *ROC curves for different noise conditions (results obtained for the Spanish SpeechDat-Car database): (a) Stopped car, engine running (12 dB). (b) High speed, good road (5 dB).*

4.2. Speech Recognition Performance

The influence of the VAD decision on the performance of a speech recognizer was also studied. The reference framework is the ETSI AURORA project for distributed speech recognition (DSR) [11] with the recognizer based on the HTK (Hidden Markov Model Toolkit) software package [12]. The influence of the VAD decision on the performance of different feature extraction schemes was studied. The first approach incorporates Wiener filtering (WF) to the base system [13] as noise suppression method. The second feature extraction algorithm that was evaluated uses Wiener filtering and non-speech frame dropping (FD).

Table 1 shows the average word accuracy (*WAcc*) for the AURORA 2 database for clean and multi-condition training/test modes. The proposed algorithm outperforms the VADs used for reference being the improvements more important when the VAD is also used for FD. The proposed VAD is the one that is closer to the "ideal" hand-labelled speech recognition performance. The improvements are more important over G.729 and AMR1 when WF and FD are applied. Table 2 shows the recognition performance averaged for the Finnish, Spanish and German SDC databases for the different training/test mismatch conditions (HM, high mismatch,

Table 1. *Recognition results for the AURORA 2 database (average WAcc for clean and multicondition training/testing).*

	Standard VADs				Other reported VAD methods				**MBQ**	Hand- labelling
	G.729	AMR1	AMR2	AFE	Woo	Li	Marzinzik	Sohn		
WF	66.19	74.97	83.37	81.57	83.64	77.43	84.02	83.89	**84.01**	84.69
WF+FD	70.32	74.29	82.89	83.29	81.09	82.11	85.23	83.80	**85.49**	86.86

Table 2. *Recognition results for the SDC databases (average WAcc for the Finnish, Spanish and German databases).*

Train/test	Standard VADs				Other reported VAD methods				**MBQ**	Base (No VAD)
	G.729	AMR1	AMR2	AFE	Sohn	Woo	Li	Marzinzik		
HM	67.93	68.59	82.58	72.53	80.52	74.95	71.80	80.52	83.98	55.08
MM	69.78	80.22	84.78	86.03	85.24	78.73	67.98	83.32	84.93	71.79
WM	88.15	93.19	94.66	94.19	94.38	91.25	71.80	93.20	94.80	92.29
Average	75.29	79.04	87.34	84.25	86.71	81.65	76.27	84.29	**87.90**	73.05

MM: medium mismatch and WM: well matched) when WF and FD are performed on the Base system. The VAD outperforms all the algorithms used for reference, yielding relevant improvements in speech recognition for both the AURORA 2 and SDC databases. Note that the SDC databases used in the AURORA 3 tasks have longer non-speech periods than the AURORA 2 database and then, more important is the effectiveness of the VAD for the speech recognition system. This fact can be clearly shown when comparing the performance of the proposed VAD to Marzinzik's VAD. The word accuracy of both VADs is quite similar for the AURORA 2 task. However the proposed VAD yields a significant performance improvement over Marzinzik's VAD for the SDC databases.

5. CONCLUSIONS

This paper presented a new VAD for improving speech detection robustness in noisy environments. The VAD is based on the estimation of the subband SNR using order statistics filters and performs and advanced and delayed detection of beginnings and word endings which leads to clear improvements in speech/pause discrimination when the SNR drops. With this and other innovations, the proposed algorithm outperformed G.729, AMR1, AMR2 and AFE standards and recently reported VAD methods in both speech/non-speech detection performance and recognition rate when considered as part of a complete speech recognition system.

6. REFERENCES

[1] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. of the International Conference on Acoustics, Speech and Signal Processing*, 1979, pp. 208–211.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 27, pp. 113–120, 1979.

[3] 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.

[4] 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.

[5] 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.

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

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

[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] 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.

[10] 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.

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

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

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