Improved Voice Activity Detection Combining Noise Reduction and Subband Divergence Measures

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

Department of Electronics and Computer Technology
University of Granada, Spain
{javierrp, segura, carmen, atv, rubio}@ugr.es

Abstract

Currently, new trends in wireless communications are demanding reliable human-machine interaction in real-life environments. However, there are obstacles inhibiting automatic speech recognition systems (ASR) working in noisy environments. The main difficulty is the degradation suffered by ASR systems due to a mismatch between training and test conditions. This paper shows an improved voice activity detector (VAD) combining noise reduction and subband divergence estimation for improving the reliability of speech recognizers operating in noisy environments. The algorithm formulates the decision rule by measuring the divergence between the subband spectral magnitude of speech and noise using the Kullback-Leibler (KL) distance on the denoised signal. Experiments demonstrate a sustained advantage over different VAD methods including standard VADs such as G.729 and AMR, which are used as a reference, recently reported algorithms, and the VADs of the advanced frontend (AFE) for distributed speech recognition (DSR).

1. Introduction

With the advent of wireless communications and the development of modern robust speech processing technology, new speech services are becoming a reality. However, many speech processing systems working in real-life scenarios encounter serious implementation barriers that affect its reliable operation. 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).

Speech/non-speech detection is an unsolved problem affecting numerous applications. The classification task is not as trivial as it appears and most of the VAD algorithms often fail in noisy environments. During the last decade different VAD methods have been proposed for several applications including mobile communication services [1], real-time speech transmission on the Internet [2] and noise reduction for digital hearing aids [3]. The detection principles are fundamentally based on the signal subband energy [4], its spectrum [5], [6], zero crossing rates (ZCR) [7], cepstral coefficients [8] and Fuzzy rules [9]. It has been shown recently that VAD robustness can be improved by measuring the Kullback-Leibler (KL) divergence between the distributions of speech and noise [10]. This paper presents several improvements over the previous work that has been shown to be very effective for noise suppression and speech recognition in noisy environments. The main contribution of this paper is the increased speech detection accuracy in high noise conditions by using a noise reduction stage previous to measuring the KL divergence on subbands. An exhaustive analysis using the popular AURORA TIdigits and SpeechDat-Car (SDC) databases provides an extensive performance evaluation and comparison to standard VADs such as ITU G.729 [7], GSM AMR [11] and the ETSI advanced front-end (AFE) [12] for distributed speech recognition (DSR), and other recently reported VADs [4, 5, 13, 14].

2. Background

In probability theory, the Kullback-Leibler (KL) divergence [15, 16] is a quantity which measures the difference between two probability distributions. Let two distributions have probability functions $p_1(x)$ and $p_2(x)$. Then the relative entropy of $p_1(x)$ with respect to $p_2(x)$, also called the KL distance, is defined by:

$$H(p_1||p_2) = \int p_1(x) \log \left(\frac{p_1(x)}{p_2(x)} \right) dx \quad (1)$$

The KL distance is not a true metric because neither the symmetry constraint nor the triangle inequality is satisfied. To make it symmetric, we in practice use the divergence distance:

$$H_s(p_1||p_2) = H(p_1||p_2) + H(p_2||p_1) \quad (2)$$

On the other hand, the computation of the KL distance is a difficult task and analytical solutions are not available except under some special circumstances. Only within certain parametric families, say the widely used Gaussian density, we have analytic expressions for the KL distance [16]. If $p_1(x)$ and $p_2(x)$ are Gaussian distributions with means μ_1 and μ_2 and standard deviations σ_1 and σ_2 , respectively, the KL divergence can be easily computed by:

$$H(p_1||p_2) = \frac{1}{2} \left[\log \left(\frac{\sigma_2^2}{\sigma_1^2} \right) - 1 + \frac{\sigma_1^2}{\sigma_2^2} + \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2} \right] \quad (3)$$

Voice activity detection usually requires evaluating measures between probabilistic distributions, among which the KL divergence distance is of great interest.

3. Improved voice activity detection algorithm

Several improvements were considered for the proposed VAD. First, a noise reduction block enables reformulating the VAD decision on a denoised signal. Second, the KL divergence is measured on the optimal number of subbands and not 23 Mel-scaled subbands.

3.1. Noise reduction block

The noisy speech signal $x(n)$ is decomposed into 25-ms frames with a 10-ms window shift. Let $X(m, l)$ be the spectrum magnitude for the m-th band ($m = 0, 1, ..., NFFT-1$) at frame l. The design of the noise reduction block is based on Wiener filter (WF) theory being its attenuation dependent on the signal-to-noise ratio (SNR) of the processed signal. The noise reduction block is shown in Fig. 1. The VAD decision is taken on the de-noised signal using the

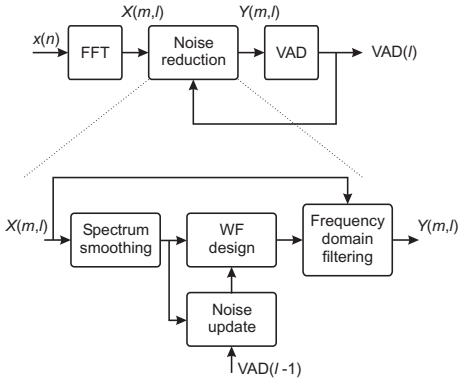


Figure 1: *Previous denoising stage for voice activity detection.*

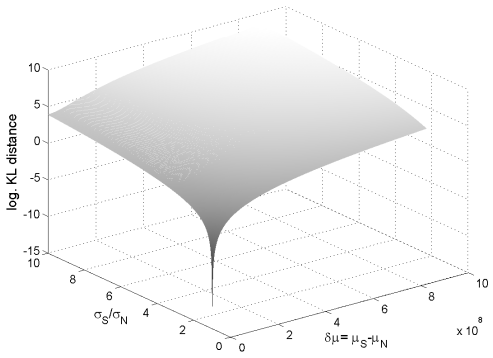


Figure 2: *KL distance as a function $\delta\mu$ of and $\delta\sigma$.*

KL divergence between speech and noise on the resulting subband energies. Noise reduction consists of:

i) *Spectrum smoothing*. The power spectrum is averaged over two consecutive frames and two adjacent spectral bands.

ii) *Noise estimation*. The noise spectrum $N_e(m,l)$ is updated during non-speech periods by means of a 1^{st} order IIR filter on the smoothed spectrum $X_s(m,l)$, that is, $N_e(m,l) = \lambda N_e(m,l-1) + (1-\lambda)X_s(m,l)$ where $\lambda = 0.99$.

iii) *WF design*. First, the clean signal $S(m)$ is estimated by spectral subtraction:

$$S(m,l) = \gamma S'(m,l) + (1-\gamma)\max(X_s(m,l) - N_e(m,l), 0) \quad (4)$$

where γ= 0.98. Then, the WF is designed as:

$$H(m,l) = \frac{\eta(m,l)}{1+\eta(m,l)}; \quad \eta(m,l) = \max\left[\frac{S(m,l)}{N_e(m)}, \eta_{\min}\right] \quad (5)$$

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

$$S'(m,l) = H(m,l)X(m,l) \quad (6)$$

The filter $H(m,l)$ is smoothed in order to eliminate rapid changes between neighbor frequencies that may often cause musical noise. Thus, the variance of the residual noise is reduced and consequently, the robustness when detecting non-speech is enhanced. The smoothing is performed by truncating the impulsive response of the corresponding causal FIR filter to 17 taps using a Hanning window.

iv) *Frequency domain filtering*. The smoother filter $\hat{H}$ is applied in the frequency domain to obtain the de-noised signal $Y(m,l) = \hat{H}(m,l)X(m,l)$.

3.2. Subband Kullback-Leibler divergence measure

Once the input speech has been de-noised, the subband energies, $E(k,l)$, in K subbands ($k = 0, 1, ..., K-1$) are computed for each frame l by means of:

$$E(k,l) = \frac{K}{NFFT}\sum_{m=m_k}^{m_{k+1}-1}|Y(m,l)|^2; \quad m_k = \left\lfloor \frac{NFFT}{2K}k \right\rfloor \quad (7)$$

Note that the process requires the noise suppression block to perform noise reduction on the block $\{X(m,l-N), X(m,l-N+1), ..., X(m,l-1), X(m,l), X(m,l+1), ..., X(m,l+N)\}$ before the subband energies $E(k,l)$ can be computed. This is carried out as follows. During the initialization process, 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 $Y(m,l+N+1)$ become available for the next iteration.

In order to evaluate the divergence between speech and silence, the VAD processes the subband energies separately by means of a (2N+1)-frame sliding window centered at the $l-th$ frame:

$$W^{(k)} = \{E(k,j)\}_{j=l-N}^{l+N} \quad (8)$$

which is subdivided as the lower and upper windows:

$$W_1^{(k)} = \{E(k,j)\}_{j=l-N}^{l-1} \qquad W_2^{(k)} = \{E(k,j)\}_{j=l+1}^{l+N} \quad (9)$$

Then, the means of the windows $W_1^{(k)}$ and $W_2^{(k)}$, $\mu_1(k)$ and $\mu_2(k)$, and their standard deviations, $\sigma_1(k)$ and $\sigma_2(k)$, are computed and on-line averaged by a 1^{st} order IIR smoothing filter:

$$\begin{aligned} \hat{\mu}_i(k) &= \alpha\hat{\mu}_i(k) + (1-\alpha)\mu_i(k) \\ \hat{\sigma}_i(k) &= \alpha\hat{\sigma}_i(k) + (1-\alpha)\sigma_i(k) \end{aligned} \qquad i = 1, 2 \quad (10)$$

where α= 0.55 is a good selection for enabling an accurate selection of word beginnings and endings. Finally, the speech/non-speech KL distance between the distributions of speech ($p_S^{(k)}(x)$) and noise ($p_N^{(k)}(x)$) is measured in K subbands using Eq. 3:

$$\rho_{S,N} = \frac{1}{2}\left[\frac{\sigma_S^2}{\sigma_N^2} + \frac{\sigma_N^2}{\sigma_S^2} - 2 + (\mu_S - \mu_N)^2\left(\frac{1}{\sigma_S^2} + \frac{1}{\sigma_N^2}\right)\right] \quad (11)$$

where the index k has been omitted for convenience. Thus, The k-band signal mean and standard deviation required by Eq. (11) are estimated using the energy window $W_2^{(k)}$ as $\mu_S(k) = \hat{\mu}_2(k)$ and $\sigma_S(k) = \hat{\sigma}_2(k)$, while noise statistics $\mu_N(k)$ and $\sigma_N(k)$ are updated during non-speech periods to track non stationary noise environments by:

$$\begin{aligned} \mu_N(k) &= \beta\mu_N(k) + (1-\beta)\min\{\hat{\mu}_1(k), \hat{\mu}_2(k)\} \\ \sigma_N(k) &= \beta\sigma_N(k) + (1-\beta)\min\{\hat{\sigma}_1(k), \hat{\sigma}_2(k)\} \end{aligned} \quad (12)$$

where β= 0.7 is the optimal value for updating the noise parameters.

Fig. 2 shows the behavior of the KL distance as a function of the deviation in the mean $\delta\mu$ and the standard deviation ratio σ_S/σ_N. It can be argued that the KL distance defined by Eq. 11 enables discriminating speech and noise taking into account both shifts in the mean and standard deviation. Finally, once the subband probability distributions for speech and noise has been modelled and their

means and spectral deviations computed, the KL "distance" is calculated through Eq. 4 and the decision rule is then formulated in terms of the mean subband KL distance:

$$\hat{\rho}_{S,N} = \frac{1}{K}\sum_{k=0}^{K-1}\rho_{S,N}(k) \tag{13}$$

which is compared to a threshold η. Thus, if the average KL distance is greater than the threshold η, the actual frame is classified as speech, otherwise it is classified as non-speech. The threshold η and the window size are made adaptive as in [17] to the noise energy E in order to select the optimum working point for different SNR conditions. On the other hand, $K = 4$ subbands was found to be the best compromise between VAD performance and complexity.

4. Performance evaluation

Several experiments were conducted to evaluate the proposed VAD. The analysis is focused on the assessment of misclassification errors and the influence of them on a speech recognition system.

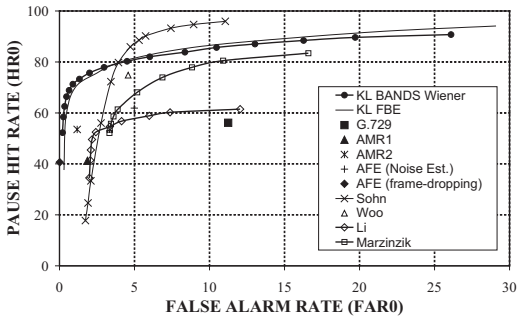
4.1. Receiver operating characteristics (ROC) curves

The ROC curves are frequently used to completely describe the VAD error rate. The AURORA subset of the original Spanish SpeechDat-Car (SDC) database [18] was used in this analysis. This database contains 4914 recordings using close-talking and distant microphones from more than 160 speakers which are categorized into three noisy conditions: quiet, low and highly noisy conditions, with 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 "actual" speech frames and "actual" speech pauses determined by hand-labelling the database on the close-talking microphone. Fig. 3 shows the ROC curves for recordings from the distant microphone and different SNR conditions. The enhanced VAD yields higher detection accuracy when compared to our previous work (KL FBE) [10] and works with lower false alarm rate and higher speech pause hit rate when compared to standards G.729 [7], AMR [11] and AFE [12](including the VADs used for noise estimation and frame-dropping) and the Sohn's [5], Woo's [14], Li's [13] and Marzinzik's [4] algorithms.

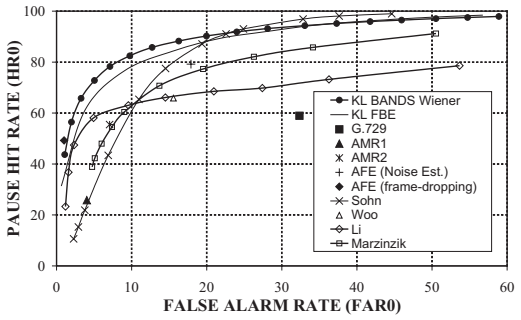
4.2. Automatic Speech Recognition Experiments

Non-efficient speech detection is an important degradation source in speech recognition systems. There are two clear motivations: *i*) Most of the speech enhancement algorithms make use of a VAD for estimating the noise statistics. Therefore, the effectiveness of the noise compensation algorithms is strongly affected by the VAD accuracy. *ii*) Frame-dropping (FD) is a frequently used technique to reduce the number of insertion errors. Since it is based on the VAD, speech frames incorrectly labelled as silence cause unrecoverable deletion errors, and silence frames incorrectly labelled as speech could increase the insertion errors.

The influence of the VAD decision on a speech recognition system was assessed by incorporating Wiener filtering (WF) and non-speech frame-dropping (FD) to the base system [19] and considering different VAD methods. 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. It is also the closest one 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



(a)



(b)

Figure 3: *ROC curves for different noise conditions: (a) Stopped car, engine running (12 dB). (b) High speed, good road (5 dB).*

averaged for the SDC databases for the different training/test mismatch conditions (HM, high mismatch, MM: medium mismatch and WM: well matched). The proposed 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 have longer non-speech periods than the AURORA 2 database and then, the effectiveness of the VAD is more important for the speech recognition performance. This fact can be clearly shown when comparing 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 for the SDC databases. Finally, in order to compare our VAD to the best available results, the VADs of the AFE standard (including both the WF and FD VADs) were replaced by the proposed VAD and the AURORA recognition experiments were conducted. The results are shown in Table 3 for the recognition experiments conducted on the SDC databases. As a result, the average word error rate is reduced from 10.83% to 10.30%.

5. Conclusion

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 KL divergence between speech and noise. Two improvements have been considered over the base system. The first of them is the selection of the optimal number of subbands. The second one reduces misclassification errors in high noisy environments by using a noise reduction stage before the KL divergence estimation. 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: *Recognition results for the AURORA 2 database (average WAcc for clean and multicondition training/testing).*

	Standard VADs				Other reported VAD methods				**KL**	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	**83.93**	84.69
WF+FD	70.32	74.29	82.89	83.29	81.09	82.11	85.23	83.80	**85.42**	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				**KL**	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	83.54	55.08
MM	69.78	80.22	84.78	86.03	85.24	78.73	67.98	83.32	84.66	71.79
WM	88.15	93.19	94.66	94.19	94.38	91.25	71.80	93.20	94.89	92.29
Average	75.29	79.04	87.34	84.25	86.71	81.65	76.27	84.29	**87.70**	73.05

Table 3: *Recognition results (word error rates).*

	AFE				
	Finnish	Spanish	German	Danish	Average
WM	3.96	3.39	4.87	6.02	**4.56**
MM	19.49	6.21	10.40	22.49	**14.65**
HM	14.77	9.23	8.70	20.39	**13.27**
Overall	**12.74**	**6.28**	**7.99**	**16.30**	**10.83**
	AFE+KL				
WM	4.29	2.94	4.65	5.87	**4.44**
MM	22.30	6.76	10.76	22.07	**15.47**
HM	10.28	7.49	8.88	17.33	**11.00**
Overall	**12.29**	**5.73**	**8.10**	**15.09**	**10.30**

6. Acknowledgements

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

7. References

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

[2] Sangwan, A., Chiranth, M. C., Jamadagni, H. S., Sah, R., Prasad, R. V., and Gaurav, V., "VAD Techniques for Real-Time Speech Transmission on the Internet", Intl. Conf. High-Speed Networks and Multimedia Comms., pp. 46–50, 2002.

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

[4] 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.

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

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

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

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

[9] 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.

[10] Ramírez, J., Segura, J. C., Benítez, M. C., de la Torre, A., Rubio, A., "A New Kullback-Leibler VAD for Speech Recognition in Noise", IEEE Signal Processing Letters, Vol. 11, No. 2, pp. 666–669, 2004.

[11] ETSI EN 301 708, Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels, 1999.

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

[13] Li, Q., Zheng, J., Tsai, A., Zhou, Q., "Robust endpoint detection and energy normalization for real-time speech and speaker recognition", IEEE Trans. Speech and Audio Processing, Vol. 10, No. 3, pp. 146–157, 2002.

[14] Woo, K., Yang, T., Park, K., Lee, C., "Robust voice activity detection algorithm for estimating noise spectrum", Electronics Letters, Vol. 36, No. 2, pp. 180–181, 2000.

[15] Aarabi, P., "Localization-based sensor validation using the Kullback-Leibler divergence", IEEE Trans. on Systems, Man, and Cybernetics, Part B, (to appear).

[16] Zhou, S. K., and Chellappa, R., "Kullback-Leibler distance between two Gaussian densities in reproducing kernel Hilter space", IEEE Intl. Symp. Information Theory, 2004.

[17] Ramírez, J., Segura, J.C., Benítez, M.C., de la Torre, A., Rubio, A., "A new voice activity detector using subband order-statistics filters for robust speech recognition", Intl. Conf. Acoustics, Speech and Signal Processing, 2004.

[18] Moreno, A., et al., "SpeechDat-Car: A Large Speech Database for Automotive Environments", II LREC Conf., 2000.

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