Noise robust model-based Voice Activity Detection

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

Departamento de Teoría de la Señal Telemática y Comunicaciones, University of Granada, Spain
[atv,javierrp,carmen,segura,luzgm,rubio]@ugr.es

Abstract

We propose a model-based VAD derived from the Vector Taylor Series (VTS) approach. A Gaussian mixture (trained with clean speech) is used in order to provide an appropriate decision rule for speech/non-speech detection. Additionally, VTS approach adapts the Gaussian mixture to noise conditions, yielding a stable performance for a wide range of SNRs. We have evaluated its ability for speech/non-speech detection and also its application for robust speech recognition. When compared to other VAD methods, the proposed VAD shows the best trade-off in speech/non-speech detection. When applied for Wiener Filtering and for frame dropping, the proposed VAD also provides the best recognition results.

Index Terms: voice activity detection (VAD), vector Taylor series approach (VTS), Gaussian mixture, Wiener filtering.

1. Introduction

Speech recognition systems are strongly affected by noise. Numerous techniques have been derived to palliate the effect of noise on the recognition performance. Most of them often require to estimate the noise statistics by means of a precise voice activity detector (VAD). The speech/non-speech classification task is not as trivial as it appears, and most of the VAD algorithms fail when the level of background noise increases. During the last decade, numerous researchers have developed different strategies for detecting speech on a noisy signal [1, 2, 3, 4] with special attention paid to the derivation and study of noise robust features and decision rules.

In this paper we propose a model-based VAD algorithm. The speech non speech decision is based on the Vector Taylor Series (VTS) approach [5, 6, 7, 8]. VTS approach, initially proposed as a noise compensation procedure for robust speech recognition, has been adapted for speech/non-speech classification. VTS formulation is based on a Gaussian mixture in the logarithmically scaled filter-bank-energy (log-FBE) domain. The Gaussian mixture is adapted to noise conditions and the noisy Gaussian mixture is used to compute the probability of each Gaussian given the noisy input frame. These probabilities are used to obtain an estimation of the clean frame. In the VTS-based VAD algorithm that we propose, these probabilities are used to compute the probability of the frame being speech. The VAD decision is done by comparing the probability of the frame being speech with a threshold. Two advantages are expected from this approach: On one hand, VAD relies on a Gaussian mixture model trained with clean speech, and therefore, the VAD decision is based on the speech events observed in the training database. On the other hand, the VTS approach provides a method to adapt the Gaussian mixture to the noise conditions. This way, the proposed method allow the adaptation of the VAD to noisy conditions and therefore, the performance of the VAD is expected to be stable for a wide range of SNRs.

2. VTS based VAD

2.1. Vector Taylor Series approach

The VTS approach [5, 6, 7, 8] is a noise compensation method providing a clean speech representation by removing the additive noise. This noise compensation is performed in the log-FBE domain and is based on a Gaussian mixture. It assumes that the effect of the noise can be described as an additive term in the log-FBE domain,

$$\mathbf{y}(\mathbf{x},\mathbf{n}) = \mathbf{x} + \mathbf{g}(\mathbf{x},\mathbf{n}) \tag{1}$$

where $\mathbf{x}$ and $\mathbf{y}$ are vectors in this domain representing the clean and noisy speech respectively, for a given frame, and $\mathbf{n}$ represents the additive noise affecting this frame. For the i^{th} channel, $\mathbf{g}$ is described by the equation,

$$g(i) = \log\left(1 + \exp\left(n(i) - x(i)\right)\right) \tag{2}$$

Two auxiliary functions $f(i)$ and $h(i)$ can be defined as,

$$f(i) \equiv \frac{1}{1+\exp(x(i)-n(i))} \qquad h(i) \equiv (1-f(i))f(i) \tag{3}$$

and using these definitions, we can approach $y(i)$ using a Taylor series around some values $x_0(i)$ and $n_0(i)$. Similarly, we can describe how a Gaussian pdf in the log-FBE domain is affected by additive noise using this Taylor series approach. Let us consider a Gaussian pdf representing clean speech, with mean $\mu_x(i)$ and covariance matrix $\Sigma_x(i,j)$ and let us assume a Gaussian noise process with mean $\mu_n(i)$ and covariance matrix $\Sigma_n(i,j)$. We can expand the Taylor series around $x_0(i) = \mu_x(i)$ and $n_0(i) = \mu_n(i)$. The mean and the covariance matrix of the pdf describing the noisy speech can be obtained as the expected values, $\mu_y(i) = E[y(i)]$ and $\Sigma_y(i,j) = E[(y(i)-\mu_y(i))(y(j)-\mu_y(j))]$, and can be estimated as a function of $\mu_x(i)$, $\mu_n(i)$, $\Sigma_x(i,j)$ and $\Sigma_n(i,j)$ using the Taylor series approach as,

$$\mu_y(i) \approx \mu_x(i) + g_0(i) + \frac{1}{2}h_0(i)[\Sigma_x(i,i) + \Sigma_n(i,i)] \tag{4}$$

$$\Sigma_y(i,j) \approx (1-f_0(i))(1-f_0(j))\Sigma_x(i,j) + f_0(i)f_0(j)\Sigma_n(i,j) + \frac{1}{2}h_0^2(i)(\Sigma_x(i,i)+\Sigma_n(i,i))^2\delta_{i,j} \tag{5}$$

where $g_0(i)$, $f_0(i)$ and $h_0(i)$ are evaluated for $x_0(i) = \mu_x(i)$ and $n_0(i) = \mu_n(i)$. Thus, the Taylor series approach gives a Gaussian pdf describing the noisy speech from the Gaussian pdf describing the clean speech and the Gaussian pdf describing the noise.

If the clean speech is modeled as a mixture of K Gaussian pdfs, the Vector Taylor Series approach provides an estimate of the clean speech $\hat{\mathbf{x}}$ given the observed noisy speech $\mathbf{y}$ and the statistics of the noise (μ_n and Σ_n) as,

$$\hat{\mathbf{x}} \approx \mathbf{y} - \sum_k P(k|\mathbf{y})\mathbf{g}(\mu_{x,k},\mu_n) \tag{6}$$

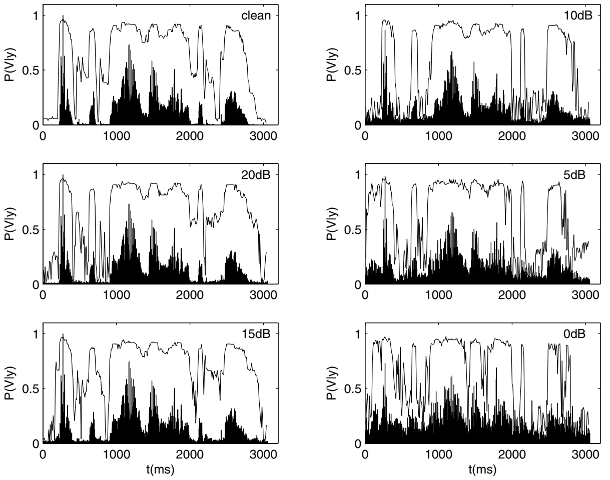

Figure 1: Probability $P(V|\mathbf{y})$ of frame $\mathbf{y}$ being speech for each frame of a sentence, evaluated at different SNRs. The signal amplitude (for positive values) has also been represented on each plot.

where $\mu_{x,k}$ is the mean of the k^{th} clean Gaussian pdf and $P(k|\mathbf{y})$ is the probability of the noisy Gaussian k generating the noisy observation $\mathbf{y}$, given by,

$$P(k|\mathbf{y}) = \frac{P(k)\mathcal{N}(\mathbf{y}, \mu_{y,k}, \Sigma_{y,k})}{\sum_{k'} P(k')\mathcal{N}(\mathbf{y}, \mu_{y,k'}, \Sigma_{y,k'})} \qquad (7)$$

where $P(k)$ is the a-priori probability of the k^{th} Gaussian and $\mathcal{N}(\mathbf{y}, \mu_{y,k}, \Sigma_{y,k})$ is the k^{th} noisy Gaussian pdf (with mean $\mu_{y,k}$ and covariance matrix $\Sigma_{y,k}$) evaluated at $\mathbf{y}$. The mean and covariance matrix of the k^{th} noisy Gaussian pdf can be estimated from the noise statistics (μ_n and Σ_n) and the k^{th} clean Gaussian pdf ($\mu_{x,k}$ and $\Sigma_{x,k}$) using equations (4) and (5). In the experiments, the parameters describing the noise statistics have been estimated using the first and the last 10 frames of each sentence (that are assumed to be silence).

2.2. Application of VTS to Voice Activity Detection

If each Gaussian k is assigned with a probability $P(V|k)$ (the probability of the k^{th} Gaussian being speech), the probability of a noisy input frame $\mathbf{y}$ being speech can be evaluated as,

$$P(V|\mathbf{y}) = \sum_{k} P(V|k)P(k|\mathbf{y}) \qquad (8)$$

where $P(k|\mathbf{y})$ is given by the VTS approach (equation (7)).

The probability $P(V|k)$ can easily be estimated for each Gaussian, since the Gaussian mixture is built from a clean speech training database. Clearly, clean Gaussians with a low mean value in the energy coefficient represent silence events, while those with a high value represent speech events. In this work we have considered the mean energy of the clean Gaussian E_k in order to estimate this probability, and a linear function between two reference energies (E_0 and E_1) has been considered,

$$P(V|k) = (E_k - E_0)/(E_1 - E_0) \text{ if } E_k \in [E_0, E_1] \qquad (9)$$

with $P(V|k)$=0 if $E_k < E_0$ and $P(V|k)$=1 if $E_k > E_1$. The parameters E_0 and E_1 have empirically been adjusted.

Fig. 1 shows the evolution over time of the evaluated $P(V|\mathbf{y})$ for a sentence extracted from AURORA-2 database [9], at different SNRs. The effectiveness of the method at low SNRs can be observed with this example. A probability greater than 0.8 is obtained for, at least, some frames at each syllable even for low SNRs (this sentence corresponds to the English digit stream 86Z1162). The decision speech/non-speech can be performed by using a threshold T. The frame $\mathbf{y}$ is labelled as speech if $P(V|\mathbf{y}) > T$, and as non-speech otherwise. Additionally, a time-in and a time-out of several frames can be considered, in order to avoid that the VAD discards those low-energy speech frames at the beginning and at the end of some syllables.

3. Experimental framework

Several experiments are commonly carried out in order to assess the performance of VAD algorithms. The analysis is normally focused on the determination of the error probabilities in different noise scenarios and SNR values [10, 4], and the influence of the VAD decision on speech processing systems [11, 12]. The experimental framework and the objective performance tests conducted to evaluate the proposed algorithm are described in this section.

3.1. Evaluation under different noise environments

First, the proposed VAD was evaluated in terms of the ability to discriminate between speech and non-speech in different noise scenarios and at different SNR levels using the AURORA-2 database [9]. This database is built from the clean TIdigits database (that consists of sequences of up to seven connected digits spoken by American English talkers) used as source speech, and a selection of eight different real-world noises that have been artificially added to the speech at SNRs of 20dB, 15dB, 10dB, 5dB, 0dB and -5dB. In the discrimination analysis, the clean TIdigits database was used to manually label each utterance as speech or non-speech on a frame by frame basis for reference. Detection performance is then assessed in terms of the speech pause hit-rate (HR0) and the speech hit-rate (HR1) defined as the fraction of all actual pause or speech frames that are correctly detected as pause or speech frames, respectively,

$$\text{HR1} = \frac{N_{1,1}}{N_1^{\text{ref.}}} \qquad \text{HR0} = \frac{N_{0,0}}{N_0^{\text{ref.}}} \qquad (10)$$

where $N_1^{\text{ref.}}$ and $N_0^{\text{ref.}}$ are the number of real speech and non-speech frames in the whole database and $N_{1,1}$ and $N_{0,0}$ are the number of speech and non-speech frames correctly classified, respectively.

Fig. 2 compares the proposed VTS-based VAD (using a threshold T=0.5) to standardized algorithms including the ITU-T G.729 [13], ETSI AMR [14] and ETSI AFE [15] and other recently reported algorithms [1, 2, 3, 4] in terms of the non-speech hit-rate (HR0) and speech hit-rate (HR1) for clean conditions and SNR levels ranging from 20 to -5 dB. Note that results for the two VADs defined in the AFE DSR standard [15] for estimating the noise spectrum in the Wiener filtering (WF) stage and non-speech frame-dropping (FD) are provided. The results shown in these figures are averaged values for the entire set of noises. It can be concluded from figure 2 that: (*i*) ITU-T G.729 VAD suffers poor speech detection accuracy with the increasing noise level while non-speech detection is good in clean conditions (85%) and poor (20%) in noisy conditions. (*ii*) ETSI AMR1 yields an conservative behavior with high speech detection accuracy for the whole range of SNR levels but very poor non-speech detection results at increasing noise levels. Although AMR1 seems to be well suited for speech detection at unfavorable noise conditions, its extremely

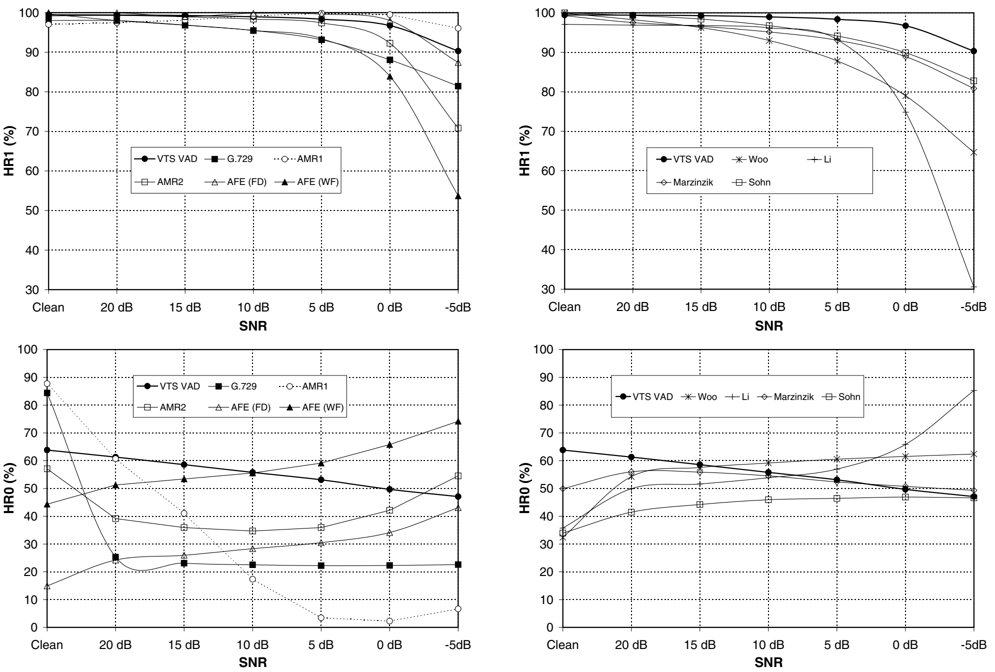

Figure 2: Speech/non-speech discrimination analysis as a function of the SNR. Results are averaged for all the noises considered in the AURORA-2 database. Speech hit-rate and non-speech hit-rate (compared to standard and other recently reported VADs).

conservative behavior degrades its non-speech detection accuracy being HR0 less than 10% below 10 dB, making it less useful in some speech processing system. (*iii*) ETSI AMR2 leads to considerable improvements over G.729 and AMR1 yielding better non-speech detection accuracy while still suffering fast degradation of the speech detection ability at unfavorable noisy conditions. (*iv*) The VAD used in the AFE standard for estimating the noise spectrum in the Wiener filtering stage is based in the full energy band and yields a poor speech detection performance with a fast decay of the speech hit-rate at low SNR values. On the other hand, the VAD used in the AFE for frame-dropping achieves a high accuracy in speech detection but moderate results in non-speech detection. (*v*) Finally, the proposed VTS-VAD yields the best compromise among the different tested VADs. It obtains a good behaviour in detecting non-speech periods as well as exhibits a slow decay in performance at unfavorable noise conditions in speech detection (90% at -5 dB).

Table 1 summarizes the advantages provided by VTS-VAD over the different VAD methods in terms of the average speech/non-speech hit-rates (over the entire range of SNR values). Thus, the proposed method with a 97.50% mean HR1 and a 55.62% mean HR0 yields the best trade-off in speech/non-speech detection when compared to all the VAD analyzed.

3.2. VAD evaluation on a robust ASR system

Although the discrimination analysis presented in the preceding section are effective for the evaluation of a given speech/non-speech discrimination algorithm, the influence of the VAD in a speech recognition system was also studied. The reference framework (Base) is the distributed speech recognition (DSR) front-end [16] proposed by the ETSI STQ working group for the evaluation of noise robust DSR feature extraction algorithms.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 as noise suppression method. The second feature extraction algorithm combines Wiener filtering and non-speech frame dropping (FD). Table 2 shows the AURORA-2 recognition results as a function of the SNR for speech recognition experiments based on the G.729, AMR, AFE, and VTS VAD algorithms. These results were averaged over the three test sets of the AURORA-2 recognition experiments. As a conclusion, the proposed VAD outperforms the standard G.729, AMR1, AMR2 and AFE VADs when used for WF and also, when the VAD removes non-speech frames.

Table 1: Speech/non-speech hit rates averaged for SNRs between clean conditions and -5 dB.

Comparison with standard VADs			
	G.729	AMR1	AMR2
HR0 (%)	31.77	31.31	42.77
HR1 (%)	93.00	98.18	93.76
	AFE (WF)	AFE (FD)	**VTS-VAD**
HR0 (%)	57.68	28.74	**55.62**
HR1 (%)	88.72	97.70	**97.50**

Comparison with other VADs					
	Sohn	Woo	Li	Marz.	**VTS-VAD**
HR0 (%)	43.66	55.40	57.03	52.69	**55.62**
HR1 (%)	94.46	88.41	83.65	93.04	**97.50**

4. Conclusions

In this paper we propose a model-based VAD derived from the Vector Taylor Series approach. The use of a Gaussian mixture (in the log-FBE domain) trained with a clean speech database provides an appropriate decision rule for speech/non-speech detec-

Table 2: Average Word Accuracy for clean training experiments (AURORA-2 database).

	Base	**Base + WF**					**Base + WF + FD**				
		G.729	AMR1	AMR2	AFE	**VTS-VAD**	G.729	AMR1	AMR2	AFE	**VTS-VAD**
Clean	99.03	98.81	98.80	98.81	98.77	98.85	98.41	97.87	98.63	98.78	98.58
20 dB	94.19	87.70	97.09	97.23	97.68	97.41	83.46	96.83	96.72	97.82	97.51
15 dB	85.41	75.23	92.05	94.61	95.19	95.07	71.76	92.03	93.76	95.28	95.37
10 dB	66.19	59.01	74.24	87.50	87.29	88.11	59.05	71.65	86.36	88.67	89.34
5 dB	39.28	40.30	44.29	71.01	66.05	70.93	43.52	40.66	70.97	71.55	73.83
0 dB	17.38	23.43	23.82	41.28	30.31	40.65	27.63	23.88	44.58	41.78	44.81
-5 dB	8.65	13.05	12.09	13.65	4.97	13.16	14.94	14.05	18.87	16.23	17.61
Average	60.49	57.13	66.30	78.33	75.30	**78.43**	57.08	65.01	78.48	79.02	**80.17**

tion. On the other hand, VTS formulation allows the adaptation of the Gaussian mixture to noise conditions, yielding a stable performance of the proposed VAD for a wide range of SNRs and noise types. The proposed VTS-VAD have been evaluated in terms of the ability to discriminate between speech and non-speech in different noise scenarios. When comparing with other standard VADs, we have found that the proposed VTS-VAD shows the best trade-off in speech/non-speech detection, with an average 97.50% mean HR1 and a 55.62% mean HR0 (averaged between clean and -5dB). With respect to the performance in speech recognition, the proposed VAD also provides the best recognition results when it is applied to the estimation of noise for Wiener Filtering and when it is applied for non-speech frame dropping.

5. Acknowledgements

This work has received research funding from the EU 6th Framework Programme, under contract number IST-2002-507943 (HIWIRE, Human Input that Works in Real Environments) and SR3-VoIP projects (TEC2004-03829/TCM) from the Spanish government. The views expressed here are those of the authors only. The Community is not liable for any use that may be made of the information contained therein.

6. References

[1] 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.

[2] 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.

[3] 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.

[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] P.J. Moreno, *Speech Recognition in Noisy Environments*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, Pensilvania, 1996.

[6] P.J. Moreno, B. Raj, and R.M. Stern, “A vector Taylor series approach for environment-independent speech recognition,” in *Proc. of ICASSP-96*, 1996, pp. 733–736.

[7] R.M. Stern, B. Raj, and P.J. Moreno, “Compensation for environmental degradation in automatic speech recognition,” in *ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, 1997, pp. 33–42.

[8] J.C. Segura, A. de la Torre, M.C. Benítez, and A.M. Peinado, “Model-based compensation of the additive noise for continuous speech recognition. Experiments using the Aurora-II database and tasks,” in *Proc. of EuroSpeech-2001*, 2001, pp. 221–224.

[9] 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.

[10] F. Beritelli, S. Casale, G. Rugeri, and S. Serrano, “Performance evaluation and comparison of G.729/AMR/Fuzzy voice activity detectors,” *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 85–88, 2002.

[11] J. Ramírez, J. C. Segura, M. C. Benítez, A. de la Torre, and A. Rubio, “Efficient voice activity detection algorithms using long-term speech information,” *Speech Communication*, vol. 42, no. 3-4, pp. 271–287, 2004.

[12] J. Ramírez, José C. Segura, C. Benítez, A. de la Torre, and A. Rubio, “An effective subband OSF-based VAD with noise reduction for robust speech recognition,” *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 6, pp. 1119–1129, 2005.

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

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

[15] ETSI, “Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms,” *ETSI ES 202 050 Recommendation*, 2002.

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