A Model based Voice Activity Detector for Noisy Environments

Kaavya Sriskandaraja[1,2], *Vidhyasaharan Sethu*[1], *Phu Ngoc Le*[1,2], *Eliathamby Ambikairajah*[1,2]

[1]School of Electrical Engineering and Telecommunications, UNSW, Australia
[2]ATP Research Laboratory, National ICT Australia (NICTA), Australia
k.sriskandaraja@student.unsw.edu.au

10.21437/Interspeech.2015-445

Abstract

This paper presents a model-based voice activity detector (VAD) aimed at operating in low signal to noise ratio conditions and non-stationary noise environments. The proposed system makes use of Gaussian mixture models trained on Mel Frequency Cepstral Coefficients extracted from noisy speech data. In addition, information from smoothed frame based log energy is used to augment the system to detect voice activity accurately. Finally, preliminary decisions made by the system are post processed to remove some false acceptances which further improves the system performance. Experimental results show that the proposed VAD significantly outperforms the system that currently produces state-of-the-art results on the QUT-NOISE-TIMIT database with relative improvements of 34.58%, 17.18% and 3.5% for high, medium and low signal to noise ratio scenarios respectively.

Index Terms: voice activity detector, low SNR, non-stationary noise, Gaussian mixture models.

1. Introduction

Voice Activity Detectors (VADs) play an integral role in a number of modern speech based systems such as speech recognitions, speech enhancement systems, speech coders, etc. Of particular interest to speech based classification systems such as speaker verification, emotion recognition and language identification systems are VADs that are able to detect if any frame of speech contains speech. While many voice activity detection algorithms have been proposed, improving VAD performance under low SNR conditions and in the presence of non-stationary noise is still an active area of research. VAD performance is of particularly significance to classification systems, where the VAD may be a bottleneck with their performance limiting overall system accuracy.

Frame based voice activity detectors (VADs) operate by extracting frame based features, typically low dimensional ones such as energy, zero crossing rate, periodicity, autocorrelation etc.[1, 2, 3], followed by a suitable rule-based or stochastic back-end that detects the presence of speech based on these features. More recent VADs have used more complex short-term features such as non-Gaussianity score [2], MFCCs [4], spectral entropy, harmonic frequency[5], higher-order statistics [6, 7] and the fusion of multiple features [8, 9].

In addition to developing suitable features, research into VADs have also considered a number of different back-ends, such as Support vector machine (SVM) [10], Classification and regression tree (CART), Gaussian likelihood ratio test (LRT) [11]. Recently, statistical model based VAD and artificial neural network based VAD in particular have attracted much attention [11, 12, 13, 14, 15, 16, 17]. Generally, VADs benefit from these back-ends based on statistical models when operating in low SNR noise conditions [11, 18] .

This paper proposes a VAD that models statistical information from frame based cepstral features with Gaussian mixture models (GMMs) and combines it with smoothed log energy to determine voicing in each frame. In addition, the proposed system post processes this decision to correct some of the false acceptances to further improve performance. The QUT-NOISE-TIMIT corpus [4] was used to evaluate the proposed system and the results are compared to a baseline system developed on this corpus [8].

2. Proposed Voice Activity Detector

The proposed VAD is similar to the unsupervised VAD described in [4] with additional modifications that adapt it to work under low SNR conditions. Functionally it combines a cepstral modelling block with a smoothed log energy thresholding block prior to decision post-processing to correct for false acceptances (refer Figure 1). Specifically, two voicing decisions for each frame are made independently based on cepstral features and log energy and only those frames where both decisions are positive are marked as containing speech. The sequence of frame based voicing decisions for each utterance are then post processed. A more detailed overview of the proposed system is presented in Figure 2.

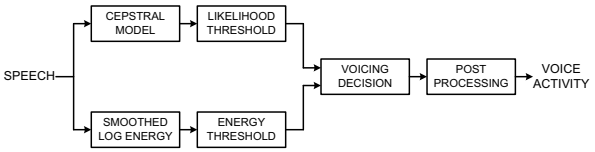


Figure 1: *Block diagram of the proposed system.*

2.1. Cepstral Modelling

The cepstral modelling aspect of the proposed VAD involves Gaussian mixture models for MFCCs estimated from speech and non-speech frames. Given a MFCC vector from a test frame, it is compared against both models and the log-likelihood ratio of speech to non-speech is estimated. This ratio is compared against a threshold estimated from appropriate training data to determine a cepstral model based decision.

This sub-system uses a 39 dimensional frame based feature vector comprising of 13 MFCC coefficients (including C_0) computed from 20ms frames using a Hamming window with 10ms overlap between frames, along with the deltas and delta-deltas. The distributions of these feature vectors for speech and non-speech frames are modelled by 16-component GMMs. The

speech to non-speech log-likelihood ratio, $L(t)$, for each frame is then given by:

$$L(t) = \log P(x_t|\lambda_S) - \log P(x_t|\lambda_{NS}) \quad (1)$$

where, $L(t)$ denotes the voicing log-likelihood ratio for frame t, x_t denotes the 39 dimensional cepstral feature vector from frame t, λ_S denotes GMM of speech frames and λ_{NS} denotes the GMM of non-speech frames.

Following this, the log-likelihood ratio contour from each utterance, $L(t)$, is smoothed using a 23-tap running mean filter in order to reduce short term variations as in [4]. The log-likelihood ratio (non-smoothed) is then compared against a threshold, θ_L, on a frame by frame basis to determine the cepstral model based voicing activity. The likelihood ratio threshold (θ_L) is estimated for each utterance from the smoothed log-like hood ratios, $\hat{L}(t)$, as follows:

$$\theta_L = \beta\hat{L}_{min} + (1-\beta)\hat{L}_{max} \quad (2)$$

Where $\hat{L}_{min} = \min(\hat{L}(t))$; $\hat{L}_{max} = \max(\hat{L}(t))$ and β is a constant that takes a value in the range $[0, 1]$.

2.2. Smoothed Log Energy

Short-term energy is an effective feature for distinguishing between speech and non-speech frames under clean conditions. However, under noisy conditions it may lead to a large number of false acceptances or misses based on the threshold value selected. In the proposed system, an energy based sub-system is used to support the cepstral modelling. Short term energy is estimated from 20ms frames with 10ms overlap and the energy contour for each utterance is smoothed using a 23-tap running mean filter to obtained the frame based smoothed log-energy, $E(t)$. The energy threshold, θ_E, against which $E(t)$ is compared to determine energy based voicing activity is estimated as:

$$\theta_E = \alpha E_{min}^{(NS)} + (1-\alpha)E_{max}^{(S)} \quad (3)$$

where $E_{min}^{(NS)}$ denotes the lowest smoothed log energy across all non-speech frames from the training data set, $E_{max}^{(S)}$ denotes the highest smoothed log energy across all speech frames from the training set and $\alpha \in [0, 1]$ is a constant.

2.3. Post Processing

The preliminary decision made by the system, based on both cepstral and energy information, for all frames in an utterance are accumulated prior to post processing to correct for systemic inaccuracies in the system. Specifically, two distinct post processing stages are applied. In the first stage, the neighbourhood of each frame marked as speech was assessed to determine how many other frames are also marked as speech. Operating on the assumption that segments of speech will not be very short, if at least 40% of the closest 2N frames (here $N = 50$ on either side) are not marked as speech, then the preliminary decision about frame under consideration is changed to non-speech. i.e., if $D(t)$ denotes the preliminary decision made about frame t, with $D(t) = 1$ denoting speech frame and $D(t) = 0$ denoting non-speech frame then the modified decision after this first stage of post processing, $\hat{D}(t)$ is given by:

$$\hat{D}(t) = \begin{cases} 0 & \text{if } \sum_{i=-N}^{N} D(t+i) < 0.8N + 1 \\ 1 & \text{Otherwise} \end{cases} \quad (4)$$

The second stage of post processing is a 'hangover scheme' that aims to address errors due to frames at the beginning and end of speech segments, at the transition of non-speech to speech and vice-versa, being classified as non-speech instead of speech. It operates by extending blocks of speech (consecutive speech frames bookended by non-speech frames) to includes previous frames spanning $300ms$ and following frames spanning $500ms$ [8].

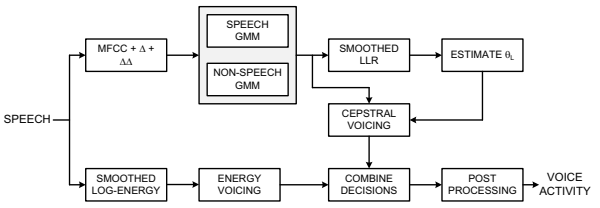


Figure 2: *Overview of the proposed system.*

3. Experimental Setup

3.1. Database

All experiments reported in this paper were carried out on the QUT-NOISE-TIMIT Corpus [19] which consists of noisy speech at various signal-to-noise ratios (SNRs). Specifically, it contains 600 hours of speech contaminated by noise recorded from 10 different locations at SNRs of 15dB, 10dB, 5dB, 0dB,-5dB and -10dB. As outlined in [9], the database categorises the 10 locations into 5 scenarios, with 2 locations corresponding to each scenario as outlined in Table 1. It can be seen that the QUT-NOISE-TIMIT corpus comprises of conditions that combine low signal-to-noise ratios with non-stationary noise, both of which can be challenging for voice activity detection systems. In this paper, SNRs of 15dB and 10dB are grouped the 'low noise' condition, 5dB and 0dB as 'medium noise' and -5dB and -10dB as 'high noise'. This grouping matches the experimental setup used to evaluate other VADs reported in literature [8, 9].

Table 1: *QUT-NOISE-TIMIT scenarios and locations.*

Scenarios	Location 1	Location 2
CAFE	Cafe	Food court
CAR	Window down	Window up
HOME	Kitchen	Living room
REVERB	Car park	Pool
STREET	City	Suburb

3.2. System Parameters and Performance Metrics

The proposed system has two hyper-parameters that should be determined, α which determines the smoothed log energy threshold (θ_E) and β which determines the cepstral model log likelihood ratio threhold (θ_L). Based on VAD performance on the training data, the value of α was chosen as 0.55, 0.45 and 0.35 for the low, medium and high noise conditions respectively. It should be noted that the running mean filter size, the number mixtures in the GMMs and the number of cepstral coefficients were similarly determined based on preliminary experiments on the training set.

The hyper-parameter β was not set to a fixed value in the experiments reported in this paper, instead it was varied uniformly between 0 and 1 to evaluate the proposed system over a range of operating points. The primary performance metrics reported are False Alarm Rate (FAR) and Miss Rate (MR), which are defined as given below. By varying β the trade-off between FAR and MR is determined and presented as Detection error trade-off (DET) curves. Both speech and non-speech Gaussian mixture models, as well as the log energy threshold, were estimated from data corresponding to all 5 noise scenarios from location 2 and all evaluation results are reported on data corresponding to the noise scenarios from location 1. It should be noted that during testing the same speech and non-speech models were used under all noise scenarios.

$$FAR = \frac{\Gamma_{NS}}{N_{NS}} \tag{5}$$

$$MR = \frac{\Gamma_{S}}{N_{S}} \tag{6}$$

where Γ_{NS} denotes the number of non-speech frames wrongly detected as speech, N_{NS} denotes the total number of non-speech frames, Γ_S denotes the number of speech frames wrongly classified as non-speech and N_S denotes the total number of speech frames.

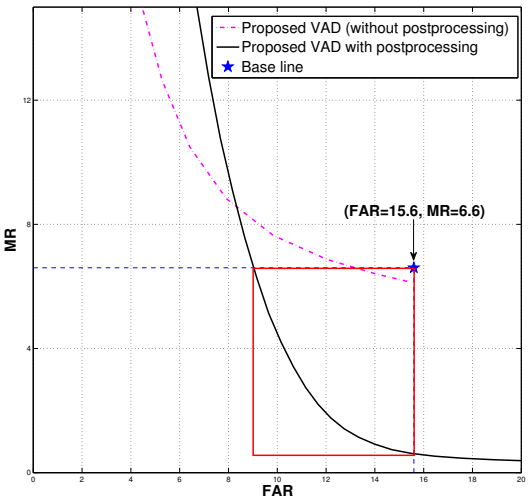
3.3. Experimental results

The performance of the proposed VAD is compared against that of the system reported in [9], which was chosen as the baseline to compare against since, to the best of the authors' knowledge, it is the best performing system evaluated on entire QUT-NOISE-TIMIT corpus. It is also evaluated with the same grouping of noise conditions and has been shown to outperform a number of other VADs including LTSD [20], Sohns VAD [11], ITU-T G.729-B [21], and the advanced front-end ETSI [22].

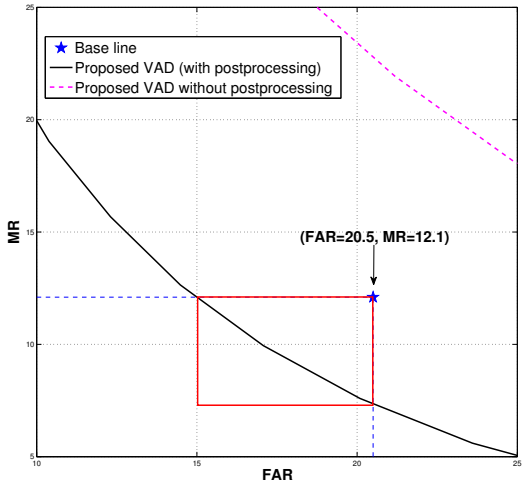
The DET curves produced by evaluating the proposed system at different operating points obtained by varying β are shown in Figures 3a, 3b and 3c for the three different noise levels - low noise ($+15dB$ and $+10dB$ SNR), medium noise ($+5dB$ and $0dB$ SNR) and high noise ($-5dB$ and $-10dB$ SNR), averaged across all 5 noise types spanning the 10 locations. The performances of the baseline system reported in [9] are indicated by a blue star marker on these DET plots. Further, the DET curves obtained by the system with and without the post-processing stage are both reported. Finally the sections of the DET curves bounded by the red rectangle indicates the operating region of the proposed VAD where it outperforms the baseline in terms of both FAR and MR.

In addition to the DET curves, the operating point (β) for the proposed VAD that resulted in a trade-off between FAR and MR that was identical to the baseline system was determined and the Half Total Error Rate (HTER) at these points for the HOME noise condition under all three noise levels are reported in Table 2. The HOME noise condition was chosen since it was one of the three highly non-stationary noise conditions available in the corpus along with CAFE and STREET. It should be noted that the HTER was evaluated as the average between the FAR and MR at the operating point.

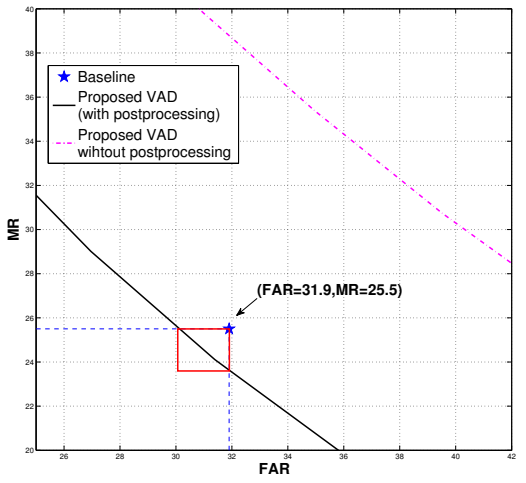
From the results presented in this section it can be seen that the proposed VAD performs significantly better than other VADs reported in the literature under low SNR conditions (down to $-10dB$) as well as in the presence of non-stationary noise. Of particular significance is the fairly large operating range within which it outperforms the baseline system in terms



(a) *DET Curve for Low Noise.*



(b) *DET Curve for Medium Noise.*



(c) *DET Curve for High Noise.*

Figure 3: *DET Curves produced by proposed VAD.*

Table 2: *%HTER for HOME noise scenario. HOME noise condition was chosen since it was one of the three highly non-stationary noise conditions available in the corpus.*

	Low noise	Medium noise	High noise
Proposed VAD	**6.46**	**9.20**	**17.21**
Baseline	20.0	23.0	30.0

of both FAR and MR. This range also provide the flexibility for the system to be tuned for varying trade-offs between FAR and MR by means of a single parameter, β.

4. Conclusions

The model based voice activity detector proposed in this paper combines information from statistical modelling of cepstral features along with information from short term smoothed log energy to determine frame based voicing decisions. A two stage post processing sub-system is proposed to improve these voicing decisions significantly, especially under low SNR conditions and in the presence of highly non-stationary noise. Our experimental results show that the proposed system significantly outperforms all other current VADs evaluated on the QUT-NOISE-TIMIT corpus. These results also suggest that the framework of proposed VAD can be expanded to incorporate other features which may lead to further improvement.

5. Acknowledgements

The authors would like to thank Prof. Sridha Sridharan, QUT, for providing the QUT-NOISE-TIMIT database and the code associated with creating the database which allowed us to evaluate and benchmark the proposed algorithm described in this paper.

6. References

[1] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell System Technical Journal*, vol. 54, no. 2, pp. 297–315, 1975.

[2] R. Chengalvarayan, "Robust energy normalization using speech/nonspeech discriminator for german connected digit recognition." in *EUROSPEECH*, vol. 99, 1999, pp. 61–64.

[3] B.-F. Wu and K.-C. Wang, "Voice activity detection based on auto-correlation function using wavelet transform and teager energy operator," *Computational Linguistics and Chinese Language Processing*, vol. 11, no. 1, pp. 87–100, 2006.

[4] M. J. Alam, P. Kenny, P. Ouellet, T. Stafylakis, and P. Dumouchel, "Supervised/unsupervised voice activity detectors for text-dependent speaker recognition on the rsr2015 corpus."

[5] L. N. Tan, B. J. Borgstrom, and A. Alwan, "Voice activity detection using harmonic frequency components in likelihood ratio test," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4466–4469.

[6] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the lpc residual domain," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 3, pp. 217–231, 2001.

[7] K. Li, M. Swamy, and M. O. Ahmad, "An improved voice activity detection using higher order statistics," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 965–974, 2005.

[8] H. Ghaemmaghami, D. Dean, S. Sridharan, and I. McCowan, "Noise robust voice activity detection using normal probability testing and time-domain histogram analysis," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4470–4473.

[9] H. Ghaemmaghami, B. J. Baker, R. J. Vogt, and S. Sridharan, "Noise robust voice activity detection using features extracted from the time-domain autocorrelation function," *Proceedings of Interspeech 2010*, 2010.

[10] J. Ramírez, P. Yélamos, J. Górriz, and J. Segura, "Svm-based speech endpoint detection using contextual speech features," *Electronics letters*, vol. 42, no. 7, pp. 426–428, 2006.

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

[12] T. V. Pham, C. T. Tang, and M. Stadtschnitzer, "Using artificial neural network for robust voice activity detection under adverse conditions," in *Computing and Communication Technologies, 2009. RIVF'09. International Conference on*. IEEE, 2009, pp. 1–8.

[13] S. Gazor and W. Zhang, "A soft voice activity detector based on a laplacian-gaussian model," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 5, pp. 498–505, 2003.

[14] D. Ying, Y. Yan, J. Dang, and F. K. Soong, "Voice activity detection based on an unsupervised learning framework," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 8, pp. 2624–2633, 2011.

[15] J. Górriz, J. Ramírez, E. W. Lang, C. G. Puntonet, and I. Turias, "Improved likelihood ratio test based voice activity detector applied to speech recognition," *Speech Communication*, vol. 52, no. 7, pp. 664–677, 2010.

[16] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 513–520.

[17] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7378–7382.

[18] M.-W. Mak and H.-B. Yu, "A study of voice activity detection techniques for nist speaker recognition evaluations," *Computer Speech & Language*, vol. 28, no. 1, pp. 295–313, 2014.

[19] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The qut-noise-timit corpus for the evaluation of voice activity detection algorithms," *Proceedings of Interspeech 2010*, 2010.

[20] J. Ramirez, J. C. Segura, C. Benitez, A. de La Torre, and A. Rubio, "Voice activity detection with noise reduction and long-term spectral divergence estimation," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 2. IEEE, 2004, pp. ii–1093.

[21] A. ITU, "silence compression scheme for g. 729 optimized for terminals conforming to recommendation v. 70," *ITU-T Recommendation G*, vol. 729, 1996.

[22] J.-Y. Li, B. Liu, R.-H. Wang, and L.-R. Dai, "A complexity reduction of etsi advanced front-end for dsr," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–61.