International Journal of Computer Applications (0975 - 8887)
Volume 182 - No.45, March 2019

Mel-Scaled Autoregressive (Mel-AR) Model based Voice Activity Detection using Likelihood Ratio Measure

M. Babul Islam
Dept. of Electrical and Electronic Engineering
University of Rajshahi, Rajshahi 6205, Bangladesh

ABSTRACT

In this paper, a Mel-scaled AR (Mel-AR) model based VAD is presented, where likelihood ratio measure is used to classify the input speech frames as speech/non-speech segments. The Mel-AR model parameters have been estimated on the linear frequency scale from the input speech signal without applying bilinear transformation. This has been done by employing a first-order all-pass filter rather than unit delay. The performance of the proposed VAD is evaluated on Aurora-2 database by measuring FAR and FRR. The equal false rate (EFR) at the crossover point is also presented as a merit of VAD. In addition, the performance of the proposed VAD in speech recognition is verified by incorporating it with a Mel-Wiener filter for MLPC based noisy speech recognition.

Keywords

VAD, Mel-AR model, Likelihood ratio, Itakura-Saito distortion, Aurora 2 database

1. INTRODUCTION

Voice activity detector (VAD) plays an important and sensitive role in many applications including robust speech recognition, digital hearing aids and discontinuous speech transmission for bandwidth reduction or distributed speech recognition over wireless and IP networks [1], [2], [3], [4]. One of the most critical problems of such applications is that the limitations of coping with the environments. Environmental noises contaminate the speech signal and change the feature parameters. As a result, the performance of these applications severely degrades in a wide variety of environmental conditions. To maintain the performance at an acceptable level a noise suppression unit along with a precise VAD is essential.

For non-stationary noises, the VAD is even more crucial since it is necessary to update constantly varying noise statistics. Therefore, a correct classification of noisy signal into speech/non-speech segments is necessary to track an accurate estimation of noise and an efficient application to a speech enhancement scheme.

Many researchers have studied different methods to develop an efficient VAD and most of them are heuristics using different speech parameters, such as, energy [5], [6], [7], zero crossing rate [2], [8], cepstral [9], LPC [10], etc. However, the algorithms based on speech features with heuristic rules have difficulty in coping with real world noises at low SNR conditions. Recently, statistical model based VAD is found to be an efficient approach to segregate speech and non-speech frames under a broad range of background noises [11], [12], [13], [14], [15], [16]. In [11], a robust VAD algorithm based on statistical likelihood ratio test (LRT) involving a single observation vector is proposed. Later, many variants of LRT have been studied to improve the performance of VAD [12], [17], [18].

In this paper, an autoregressive (AR) model [19] based VAD is proposed, where likelihood ratio (LR) measure is used to classify the input speech frames as speech/non-speech segments. The AR model is implemented on mel-scale using a first-order all-pass filter instead of unit delay.

2. GAUSSIAN MEL-SCALED AUTOREGRESSIVE MODEL

The frequency warped signal $\tilde{x}[n]$ $(n = 0, 1, .., \infty)$ obtained by the bilinear transformation [20] of a finite length windowed signal $(n = 0, 1, .., N-1)$ is defined as

$$\tilde{X}(\tilde{z}) = \sum_{n=0}^{\infty} \tilde{x}[n]\tilde{z}^{-n} = X(z) = \sum_{n=0}^{N-1} x[n]z^{-n} \qquad (1)$$

where $\tilde{z}^{-1}$ is the first-order all-pass filter,

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha \cdot z^{-1}}. \qquad (2)$$

The phase response of $\tilde{z}^{-1}$ is given by

$$\tilde{\lambda} = \lambda + 2 \cdot \tan^{-1}\left\{\frac{\alpha \sin\lambda}{1 - \alpha\cos\lambda}\right\} \qquad (3)$$

This phase function determines a frequency mapping, and $\alpha (0 < \alpha < 1)$ controls the degree of warping.

Let x be an N-dimensional random variable corresponds to N consecutive samples of the windowed signal. For an M-th order zero mean autoregressive process, x is given by

$$\tilde{e}_n = \sum_{i=0}^{M} \tilde{a}_i \tilde{x}_{n-i} \qquad (n = 0, .., \infty) \qquad (4)$$

where $\{\tilde{e}_n\}$ are Gaussian i.i.d. random variables with zero mean and unity variance, and $\{\tilde{a}_i\}$ are the Mel-scaled AR coefficients with $\tilde{a}_0 = 1$.

Now, for large N, the probability density function for x can be approximated by [19]

$$f_a(\boldsymbol{x}) \approx (2\pi)^{-N} \exp\{-\frac{1}{2}\delta(\boldsymbol{x};\tilde{\boldsymbol{a}})\} \tag{5}$$

where

$$\delta(\boldsymbol{x};\tilde{\boldsymbol{a}}) = R_{\tilde{a}}[0]\tilde{r}_x[0] + 2\sum_{i=1}^{M} R_{\tilde{a}}[i]\tilde{r}_x[i] \tag{6}$$

$R_{\tilde{a}}[i]$ is the autocorrelation function of AR coefficients and $\tilde{r}_x[i]$ is the mel-autocorrelation function [21], [22], [23] of x.

The assumption made here is that the signal x has already been properly scaled, that is, in the LPC terminology this is equivalent to normalization by the square root of average residual energy.

3. LIKELIHOOD RATIO MEASURE

The proposed VAD is based on the likelihood ratio measure between autoregressive model of noise and input speech signal. An M-th order autoregressive noise model with coefficients $\tilde{a}_0 = 1$ is created from initial 20 frames of the input speech signal. Then for any speech frame t, the mel-autocorrelation function $\tilde{r}_x[i]$ is calculated to estimate likelihood ratio between AR noise model and current speech frame as follows:

$$d_{LR} = R_{\tilde{a}}[0]\tilde{r}_{xx}[0] + 2\sum_{i=1}^{M} R_{\tilde{a}}[i]\tilde{r}_{xx}[i] - 1 \tag{7}$$

Finally, d_{LR} is compared with a threshold value η. For $d_{LR} < \eta$, the frame is detected as noise, otherwise, speech frame.

When a frame t is detected as noise, the estimated mel-autocorrelation function of noise $\hat{\tilde{r}}_n[i,t]$ is updated by accumulating $\tilde{r}_x[i,t]$ as follows:

$$\hat{\tilde{r}}_n[i,t] = \begin{cases} \beta\hat{\tilde{r}}_n[i,t_p] + (1-\beta)\tilde{r}_x[i,t]; \\ \qquad\qquad \text{if frame } t \text{ is silence} \\ \hat{\tilde{r}}_n[i,t_p]; \\ \qquad\qquad \text{if frame } t \text{ is speech} \end{cases} \tag{8}$$

where t_p is the previous noise frame and β is the forgetting factor of value $0 < \beta < 1$.

Though the proposed VAD is based on the likelihood ratio measure, it is also possible to implement the VAD based on Itakura-Saito distortion measure [24]. Itakura-Saito distortion measure d_{IS} between AR noise model and input speech frame is given by

$$d_{IS} = \frac{1}{\sigma_{en}^2}\delta(\boldsymbol{x};\tilde{\boldsymbol{a}}) + \log\frac{\sigma_{en}^2}{\sigma_{ex}^2} - 1 \tag{9}$$

where σ_{en}^2 and σ_{ex}^2 are the residual energies of the estimated noise and current frame, respectively, and $\delta(\boldsymbol{x};\tilde{\boldsymbol{a}})$ is given by Eq. (6).

4. EXPERIMENTAL SETUP

The proposed VAD was evaluated on test set A in Aurora 2 database [25]. The Aurora 2 database is a subset of TI digits database [26] contaminated by additive noises and channel effects. The order of AR model was set to 10 and the window length was 40 ms with 10 ms frame period. The value of forgetting factor was set to 0.96.

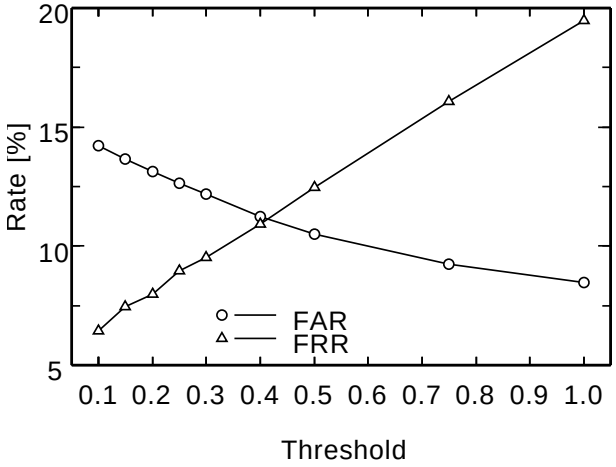

Fig. 1. False alarm and false rejection rate as a function of threshold.

5. PERFORMANCE EVALUATION

Usually two measures are used to examine the VAD performance. One is frame based false alarm rate (FAR) and the other one is frame based false rejection rate (FRR). As reference the corresponding clean speech files are labeled as speech/nonspeech frames using an energy based VAD. Because for clean speech the energy based VAD can properly discriminate speech and silence.

As the threshold factor η is used for detecting input frames as speech or noise, the effect of threshold factor on FAR and FRR is examined and the result is presented in Figure 1. Here FAR and FRR are calculated by using all the speech files for the entire set of noises (subway, babble, car and exhibition) in test set A for 5 dB SNR. The experiment was carried out for the threshold factor of 0.0 to 1.0. As shown in Figure 1, the proposed VAD keeps a steady FAR and FRR with increasing threshold factor. It is also observed that the FAR has a decreasing trend with increasing threshold factor. On the other hand, reverse characteristic is seen for FRR. The higher value of FRR means the most of the noise frames are detected as speech, on the contrary, the higher value of FAR means the most of the speech frames are detected as noise. Hence, there should be a trade-off between FAR and FRR for better estimation of noise. It has been found that the crossover point is obtained at the value of threshold factor $\eta = 0.41$, and the equal false rate (EFR) at this point is around 11.2%.

The EFR at the crossover point both for Itakura-Saito (IS) distortion and likelihood ratio (LR) measure as a function of window length has also been examined and presented in Figure 2. It has been found that longer window length gives lower EFR both for IS and LR based VAD. Consequently, the proposed system uses 40 ms window length for VAD. Though the EFR for IS based VAD is much lower than that of LR based VAD, the recognition result for test set A of Aurora 2 database shows that LR based VAD obtains slightly better result, which is presented in Figure 4.

To find the optimum threshold value, a number of recognition experiments were carried out with different threshold values under the conditions given in Table 1. The threshold factor $\eta = 0.0$ means the noise model is not adaptive and it is created from the initial 20 frames of the speech signal. The larger threshold affects the esti-

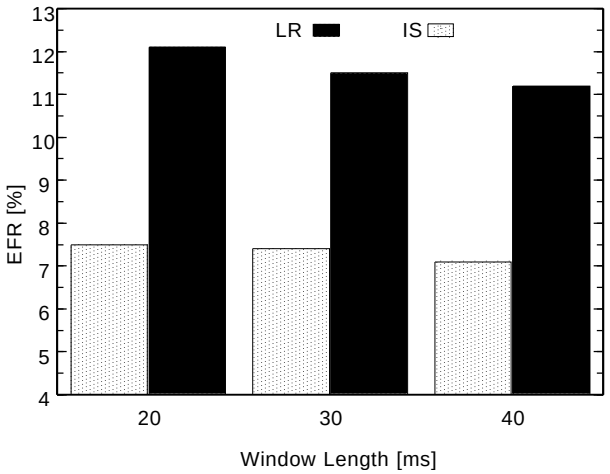

Fig. 2. False rate at crossover point as a function of window length both for IS and LR based VAD.

mated noise model and changes the model into speech like model. Because of higher value of threshold factor most of the frames are detected as noise frames. As shown in Figure 3, the highest recognition accuracy is obtained at the values of threshold factor around 0.1 to 0.15. In the final recognition experiment the threshold value was set to 0.11.

Table 1. Analysis conditions for recognition experiment.

Front-end	Feature: MLPC Analysis order: 12 Window length: 20 ms Frame shift: 10 ms Feature vector: 14 cep + 14 Δ
Enhancement	Filter: Mel-Wiener Order: 3 VAD: Proposed Cepstral processing: Blind equalization
Back-end	HMM

In Figure 5, FAR and FRR are presented as a function of SNR. The false alarm and false rejection rate are calculated by averaging over all noises for SNRs 15 to 0 dB with 5 dB interval. At high SNR conditions both the FAR and FRR are almost constant. This means that at a certain level of SNR, the performance of the proposed VAD is almost unchanged with increasing value of SNR. At SNR 5 dB an unexpected result is obtained.

6. CONCLUSION

This paper presents an autoregressive model based VAD and its application to the robust speech recognition. The autoregressive model is efficiently implemented on mel-scale. The likelihood ratio measure is used to segregate speech and non-speech frames. The performance of the proposed VAD is evaluated on Aurora 2 database. The FAR, FRR and EFR are presented as the merit of VAD. The recognition accuracy for MLPC based front-end with

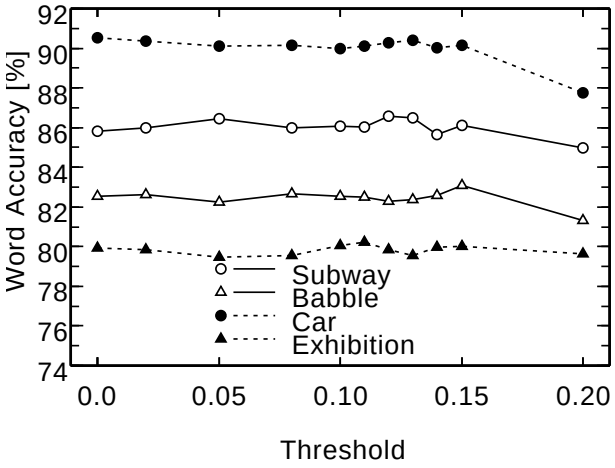

Fig. 3. Recognition accuracy as a function of threshold.

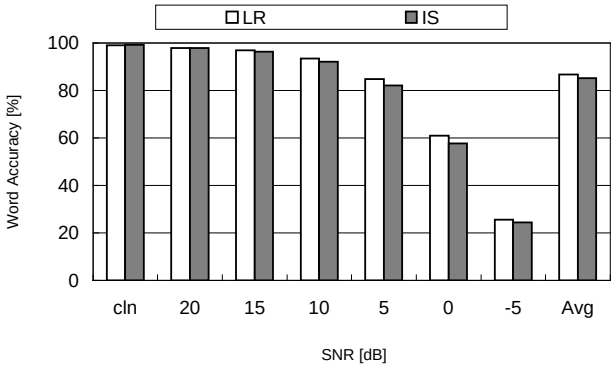

Fig. 4. Recognition accuracy for LR and IS based VAD using MLPC based front-end with Mel-Wiener filter.

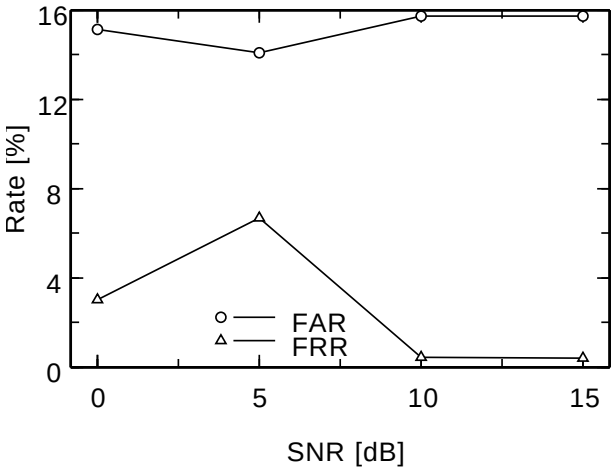

Fig. 5. False alarm and false rejection rate as a function of SNR.

mel-Wiener filter and proposed VAD is found to be 87.04% for test set A.

7. REFERENCES

[1] J. Ramirez and et. al. 2004 A new KullbackLeibler VAD for speech recognition in noise. IEEE Signal Processing Letters, 11(2): 266-269.

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

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

[4] ETSI. 2007. Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms. ETSI ES 202 050 v1.1.5.

[5] Asgari, M. 2008. Voice Activity Detection Using Entropy in Spectrum Domain. Telecommunication Networks and Applications Conference, 407-410.

[6] Evanglelopulos, G. and Maragos, P. 2006. Multiband modulation energy tracking for noisy speech detection. IEEE Trans. Audio, Speech and Lang. Process, 14(6), 2024-2038.

[7] Padrell, J., Macho, D. and Nadeu, J. 2005. Robust speech activity detection using LDA applied to FF parameters. Proceedings ICASSP'05, 1: 557-560.

[8] Bachu, R. G. et al. 2010. Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy. Advanced Techniques in Computing Sciences and Software Engineering, K. Elleithy, Ed., ed: Springer Netherlands, 279-282.

[9] Fukuda, T. Ichikawa, O. and Nishimura, M. 2010. Improved voice activity detection using static harmonic features. Proceeding ICASSP'10, 4482-4485.

[10] Li, K., et al. 2005. An improved voice activity detection using higher order statistics. IEEE Trans. Speech and Audio Process, 13(5): 965-974.

[11] Sohn, J. et al. 1999. A statistical model-based voice activity detection. IEEE Signal Process. Letters, 16(1): 1-3.

[12] Cho, Y. D. et al. 2001. Improved voice activity detection based on a Smoothed statistical likelihood ratio. Proceedings ICASSP'01, 2: 737-740.

[13] Gorriz, J. M. et al. 2008. Jointly Gaussian PDF-Based Likelihood Ratio Test for Voice Activity Detection. IEEE Trans. On Audio, Speech and Lang. Process, 16(8): 1565-1578.

[14] Fujimoto, M. et al. 2007. Noise Robust Voice Activity Detection based on Statistical Model and Parallel Non-linear Kalman Filtering. Proceedings ICASSP'07, 4: 797-800.

[15] Bao, X. and Zhu, J. 2012. A novel voice activity detection based on phoneme recognition using statistical model, EURASIP Journal on Audio, Speech, and Music Processing, 2012(1): 1-10.

[16] Tan, L. N.et al. 2010. Voice activity detection using harmonic frequency components in likelihood ratio test, ICASSP'10, 4466-4469.

[17] Ramirez, J. et al. 2007. Improved Voice Activity Detection Using Contextual Multiple Hypothesis Testing for Robust Speech Recognition. IEEE transactions on audio, speech and language processing, 15(8): 2177-2189.

[18] Gorriz, J. M. et al. 2005. An improved MO-LRT VAD based on a bispectra Gaussian model. Electronics Letters, 41(15): 877-879.

[19] Juang, B. 1984. On the hidden Markov model and dynamic time warping for speech recognition - a unified view. AT&T Bell Lab. Tec. Journal, 63(7): 1213-1243.

[20] Oppenheim, A. V. and Johnson, D. H. 1972. Discrete representation of signals. IEEE Proc., 60(6): 681-691.

[21] Strube, H. W. 1980. Linear prediction on a warped frequency scale. J. Acoust. Soc. America, 68(4): 1071-1076.

[22] Matsumoto, H., et al. 1998. An efficient Mel-LPC analysis method for speech recognition. Proc. of ICSLP'98: 1051-1054.

[23] Islam, M. B., et al. 2007. Mel-Wiener filter for Mel-LPC based speech recognition. IEICE Transactions on Information and Systems, E90-D (6): 935-942.

[24] Itakura, F. and Saito, S. 1968. Analysis synthesis telephony based on the Maximum Likelihood Method. Proc. of 6th International Congress on Acoustic, C17-C20.

[25] Hirsch, H. G. and Pearce, D. 2000. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. Proc. ISCA ITRW ASR 2000: 181-188.

[26] Leonard, R. G. 1984. A database for speaker independent digit recognition. ICASSP'84, 3: 42.11.1-42.11.4.