Voice Activity Detection using Temporal Characteristics of Autocorrelation Lag and Maximum Spectral Amplitude in Sub-bands

Sivanand Achanta[1], **Nivedita Chennupati**[1], **Vishala Pannala**[1], **Mansi Rankawat**[2], **Kishore Prahallad**[1]
[1]Speech and Vision Lab, Language Technologies Research Center,IIIT - Hyderabad, India.
[2]Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India
{sivanand.a,nivedita.chennupati,p.viahala}@research.iiit.ac.in,
mansirankawat19@gmail.com, kishore@iiit.ac.in

Abstract

A robust voice activity detection (VAD) is a prerequisite for many speech based applications like speech recognition. We investigated two VAD techniques that use time domain and frequency domain characteristics of speech signal. The temporal characteristic of the autocorrelation lag is able to discriminate speech and nonspeech regions. In the frequency domain, peak value of the magnitude spectrum in different sub-bands is used for VAD.

Performance of the proposed methods are evaluated on TIMIT database with noises from NOISEX-92 database at various signal-to-noise ratio (SNR) levels. From the experimental results, it is observed that VAD based on autocorrelation lag is working consistently better than the maximum peak value of the autocorrelation function based method. However, it performs inferior compared to our second approach and AMR-VAD2. Our second approach i.e., VAD based on maximum spectral amplitude in sub-bands outperforms AMR-VAD2 and Sohn VAD for some noise conditions. Moreover, it is shown that a threshold independent of noises and their levels can be selected in the proposed method.

1 Introduction

Voice activity detection (VAD) aims at separating the background noise and speech. VAD plays an important preprocessing role in applications like automatic speech recognition (Karray and Martin, 2003), speaker verification (Kinnunen and Rajan, 2013), wireless communications (Beritelli et al., 1998), speech enhancement for hearing aids (Itoh and Mizushima, 1997), etc. So, there has been growing interest for developing a robust VAD in low signal-to-noise ratio (SNR) conditions.

Approaches to VAD can be broadly classified as model-based and non-model based (signal processing) methods. One of the recent model-based approaches is based on using non-negative sparse coding (Teng and Jia, 2013). In this, a dictionary is trained for speech and noise separately and are concatenated to form a global dictionary. The noisy signals are represented as linear combination of elements of global dictionary. One inherent drawback of this technique is that it assumes noise during the test time to be known apriori.

In addition, there are also statistical model-based VADs (Sohn et al., 1999) (Ramirez et al., 2005) (Tan et al., 2010). Here, typically the noisy speech complex spectrum is assumed to follow a distribution like Gaussian and the parameters are estimated using various methods. This is followed by a likelihood ratio test on each frame to declare the signal frame to be speech absent or speech present. Improvements to incorporate continuity (Ramirez et al., 2005) and robustness (Tan et al., 2010) have also been proposed. Most of these techniques assume the noise statistics like variance to be known apriori. In general, these techniques perform poorly in low SNR conditions (You et al., 2012).

On the other hand, there are signal processing based approaches like using long-term signal variability (Ghosh et al., 2011), spectral flux (Sadjadi and Hansen, 2013), time-domain autocorrelation function (Ghaemmaghami et al., 2010), sub-band order statistic filters (Ramirez et al., 2004) to the VAD problem. These primarily involve ex-

D S Sharma, R Sangal and J D Pawar. Proc. of the 11th Intl. Conference on Natural Language Processing, pages 48–55,
Goa, India. December 2014.

tracting a feature which is specific to speech and robust to various noises. For example, method based on time autocorrelation function proposed in (Ghaemmaghami et al., 2010), uses maximum peak of autocorrelation function (at non-zero lag) as the feature along with quasi periodicity property of speech to improve the robustness of VAD. In our time domain approach, we compare the performance of VAD using maximum peak of autocorrelation function (at non-zero lag) as a feature against the corresponding lag of autocorrelation function. The method using maximum peak of autocorrelation function (at non-zero lag) is referred to as ACF-MAX and that using corresponding lag is referred to as ACF-LAG hereafter. In frequency domain, the maximum amplitude of magnitude spectrum in sub-bands is used as a feature for VAD, we refer to this method as MSA-SB. While ACF-LAG method can be looked upon as an excitation based method, the MSA-SB can be accounted as a system based technique. Our techniques use speech production based features and are expected to be robust to a wide variety of noise conditions.

Our contributions in this paper are, investigating robustness of autocorrelation lag over peak method, proposing the use of maximum spectral amplitudes in speech specific sub-bands and combining these contours along with mean, variance normalizations to get a threshold independent of noises and their dBs.

Rest of the paper is organized as follows. The database and evaluation metrics used are described in Section 2. The detailed description of time domain approach is given in Section 3. Section 4 discusses the frequency domain technique. Conclusions follow in Section 5.

2 Database and Evaluation Metrics

The test signals are created by taking clean speech signals from TIMIT (tim, 1993) corpus and synthetically adding noise from the NOISEX-92 (Varga and Steeneken, 1993) corpus. Around 80 signals from TIMIT corpus sampled at 16000 Hz are taken. 10 signals from each of eight dialects with 7 male and 3 female sentences are randomly selected. Every signal is appended with approximately 2 sec silence before and after the speech signal and then noise is added to it at desired SNR. Seven different noises are used from NOISEX-92 database and SNRs at -10dB, -5dB, 0dB and 5dB are considered. Approximately, each test signal has 40 % of noisy speech part and 60 % of noise part. The ground truth is generated by considering the appended silence along with labels of ‘ h# ’, ‘ pau ’ and ‘ epi ’ in the TIMIT phone file as nonspeech and the other regions as speech. False alarm rate (% FAR) and miss rate (% MR) are used as evaluation metrics, and are given by,

$$\%\text{FAR} = \left(\frac{\text{nonspeech samples detected as speech}}{\text{total number of nonspeech samples}} \right) \times 100$$

$$\%\text{MR} = \left(\frac{\text{speech samples detected as nonspeech}}{\text{total number of speech samples}} \right) \times 100$$

The half total error rate (HTER) (Ghaemmaghami et al., 2010) is computed as the mean of FAR and MR. For a good VAD algorithm, FAR, MR and HTER must be as low as possible.

3 The Time Domain Method

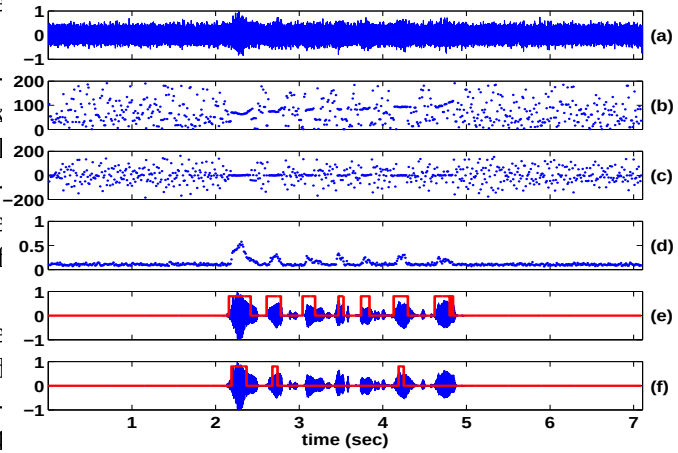

Figure 1: *Illustration of ACF-LAG and ACF-MAX methods; (a) Noisy speech signal (white noise at -10 dB), (b) Lag at the maximum in ACF plot, (c) Difference of (b), (d) Maximum peak of normalized ACF, (e) VAD from ACF-LAG method (displayed on clean speech signal for reference), (f) VAD from ACF-MAX method (displayed on clean speech signal for reference)*

The time domain autocorrelation function has been used in the past for many basic speech processing tasks like pitch extraction (de Cheveign and Kawahara, 2002). These methods exploit two key features associated with the autocorrelation function, one is the lag of the maximum peak which is usually used to compute F_0 and the other

is the amplitude of maximum peak which is used to decide whether a speech frame is voiced or unvoiced. The maximum amplitude of autocorrelation function (ACF-MAX) is not a robust feature in low SNR conditions. So, in (Ghaemmaghami et al., 2010), along with ACF-MAX, the quasi periodicity property of speech is incorporated as a feature by using the cross-correlation to take the VAD decision. To exploit quasi-periodicity of speech, we propose to use the lag of the autocorrelation function (ACF-LAG) as a feature for VAD. The basis for our method comes from the observation that the pitch period of speech signals is locally stationary and varies smoothly in voiced regions (e.g., Fig. 1(b) region around 2-2.5 sec) where as in noise or unvoiced regions the lag varies erratically (e.g., Fig. 1(b) region around 0-1 sec). It is this speech specific feature which is exploited here to detect speech and noisy regions in a given signal. To the best of authors knowledge, pitch period or lag has not been solely used for VAD previously. Hence, ACF-LAG performance is analysed for VAD in this section.

In our method, the input speech is segmented into frames with frame size of 20 ms and shift of 10 ms. Let $x_p[n]$ be the p^{th} signal frame, the normalised autocorrelation function for the frame is computed as,

$$R_p[l] = \frac{\sum_{n=0}^{L-l-1} x_p[n]x_p[n+l]}{\sum_{n=0}^{L-1} x_p[n]x_p[n]} \quad (1)$$

where l is the autocorrelation lag and L is the length of the signal frame. Usually l is limited between 2 ms and 20 ms because any value of pitch outside this range is considered to be spurious.

$$V(p) = \max\ R_p(l) \quad (2)$$

$$I(p) = \underset{l}{\operatorname{argmax}}\ R_p(l) \quad (3)$$

where $V(p)$ is the peak of autocorrelation function at non-zero lag and $I(p)$ is the corresponding lag at which the peak occurs per frame.

In ACF-MAX method, peak of autocorrelation function (eq. 2) is thresholded to get the VAD decision. In ACF-LAG method, VAD decision is made using the lag (eq. 3) corresponding to maximum of autocorrelation function. $I(p)$ is plotted in Fig. 1(b). From the plot, it can be seen that for unvoiced/noise regions the values of index vary randomly where as in voiced regions, it varies smoothly. This characteristic of the contour is used to detect voiced and unvoiced/noise regions in speech. The difference operation on contour, will give its slope and slope should be minimal when the contour is slowly varying. VAD decision is taken by setting a threshold on differenced vector. Fig. 1 (e) and (f), shows VAD decisions from ACF-LAG and ACF-MAX methods respectively. It can be seen that ACF-LAG method performs better than ACF-MAX method.

Table 1: *FAR and MR for various noises in different SNRs for ACF-LAG and ACF-MAX methods*

White Noise	**5 dB**		**0 dB**		**-5 dB**		**-10 dB**	
	MR %	FAR %	MR %	FAR %	MR %	FAR %	MR %	FAR %
ACF-LAG	**52.90**	0.05	**58.61**	0.04	**67.64**	0.04	**81.96**	0.01
ACF-MAX	71.21	0.00	82.82	0.00	93.79	0.00	99.24	0.00

Pink Noise	**5 dB**		**0 dB**		**-5 dB**		**-10 dB**	
	MR %	FAR %	MR %	FAR %	MR %	FAR %	MR %	FAR %
ACF-LAG	**57.90**	0.07	**65.64**	0.04	**78.22**	0.04	**91.95**	0.09
ACF-MAX	69.71	0.00	81.44	0.00	92.62	0.00	98.89	0.00

HFchannel Noise	**5 dB**		**0 dB**		**-5 dB**		**-10 dB**	
	MR %	FAR %	MR %	FAR %	MR %	FAR %	MR %	FAR %
ACF-LAG	**55.49**	0.09	**63.38**	0.09	**75.44**	0.16	**89.96**	0.09
ACF-MAX	70.40	0.00	82.01	0.00	93.70	0.00	98.94	0.00

Factory1 Noise	**5 dB**		**0 dB**		**-5 dB**		**-10 dB**	
	MR %	FAR %	MR %	FAR %	MR %	FAR %	MR %	FAR %
ACF-LAG	**55.52**	3.29	**63.46**	4.70	**76.03**	3.83	**88.07**	3.62
ACF-MAX	69.96	0.02	81.25	0.01	92.03	0.00	98.00	0.00

Buccaneer1 Noise	**5 dB**		**0 dB**		**-5 dB**		**-10 dB**	
	MR %	FAR %	MR %	FAR %	MR %	FAR %	MR %	FAR %
ACF-LAG	**57.17**	0.21	**65.16**	0.26	**78.23**	0.19	**90.99**	0.28
ACF-MAX	70.73	0.00	83.04	0.00	94.25	0.00	99.29	0.00

Volvo Noise	**5 dB**		**0 dB**		**-5 dB**		**-10 dB**	
	MR %	FAR %	MR %	FAR %	MR %	FAR %	MR %	FAR %
ACF-LAG	**56.61**	0.07	**63.99**	0.04	**75.32**	0.01	**87.07**	0.00
ACF-MAX	63.70	0.01	72.07	0.01	83.25	0.00	92.36	0.00

Babble Noise	**5 dB**		**0 dB**		**-5 dB**		**-10 dB**	
	MR %	FAR %	MR %	FAR %	MR %	FAR %	MR %	FAR %
ACF-LAG	**51.25**	19.28	**59.45**	16.71	**70.41**	17.81	**77.93**	17.27
ACF-MAX	66.33	3.58	77.52	3.07	89.15	3.93	95.17	3.51

3.1 Results

Table 1 reports MR and FAR of ACF-LAG and ACF-MAX methods. FAR is low for both the methods across all the noises at different SNRs. This implies that rejection of nonspeech by both the algorithms is equally good. It can also be observed from the Table 1 that ACF-LAG method has relatively lower MR than the ACF-MAX method. Hence, our hypothesis that lag of the autocorrelation function at the maximum is a robust feature compared to the peak value itself is evident. The MR is high in both the methods indicating that actual speech is missed in most of the cases. This is due to the fact that proposed methods work only for voiced regions but ground

truth includes both voiced and unvoiced regions as speech. Thus both the techniques are far from being useful as a practical VAD and hence we explore the frequency domain approach.

4 The Frequency Domain Method

The resonances of the vocal tract are high energy regions in the spectrum and are hence expected to be robust to noisy conditions. Due to inherent constraints in the human speech production mechanism, the variation of spectrum is slow as compared to noisy regions. This fact has been used in the literature for VAD, by utilizing feature such as spectral flux. However, our technique differs from all the previous techniques by making use of maximum of the magnitude spectrum alone as the feature. The maximum in magnitude spectrum corresponds to the strength of a resonance of vocal tract in speech regions and is used as a feature to distinguish speech from nonspeech.

The given noisy signal is first segmented into frames with frame size of 25 ms and hop of 5 ms. Each frame is windowed with a hamming window. The discrete Fourier transform (DFT) for p^{th} frame of the signal is computed as,

$$X_p[k] = \sum_{n=0}^{N-1} x_p[n] e^{-\frac{j2\pi kn}{N}} \qquad (4)$$

where N is the number of DFT points and k ranges from $0, \cdots, N-1$. N is set to 2048 in our experiments. Then the maximum of the magnitude part of the complex spectrum for each frame is the desired spectral feature.

$$M(p) = \max |X_p(k)|; \quad k = 0, 1, \cdots, N-1 \qquad (5)$$

In Fig. 2, noisy signal (signal corrupted with white noise at -5 dB) is shown in (a) and the corresponding maximum of the DFT spectrum extracted per frame is plotted in (b). It is observed that in the noise part, there is a high frequency ripple (e.g., Fig. 2(b) region around 0-1 sec) and in the speech region the variation of maximum over time is slow and smooth (e.g., Fig. 2(b) region around 2-3 sec). So, an FIR filter is used for low-pass filtering to remove the ripple. The low-pass filtered version of the maximum contour is plotted in (c) which is then thresholded to take the VAD decision.

While this method works for white noise, it fails for few noises like pink and volvo. As can be seen from the Fig. 3 (b) and (c), for pink noise (at -5 dB SNR), passing the maximum contour through the low-pass filter, even the noisy region has a slowly varying maximum amplitude. This is because of the high concentration of low frequency energy in pink noise.

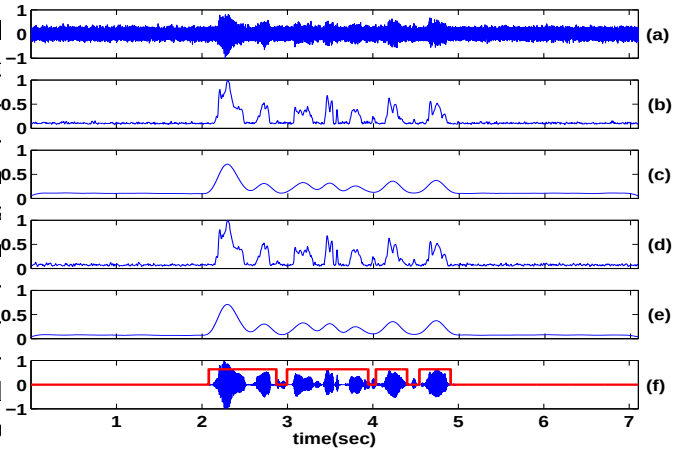

Figure 2: *The maximum contours of the DFT spectrum in white noise at -5 dB; (a) Noisy signal, (b) Maximum amplitude in the magnitude spectrum, (c) Low-pass filtered signal of (b), (d) Maximum amplitude in the resonance 1 sub-band of magnitude spectrum, (e) Low-pass filtered signal of (d), (f) VAD from MSA-SB method (displayed on clean speech signal for reference)*

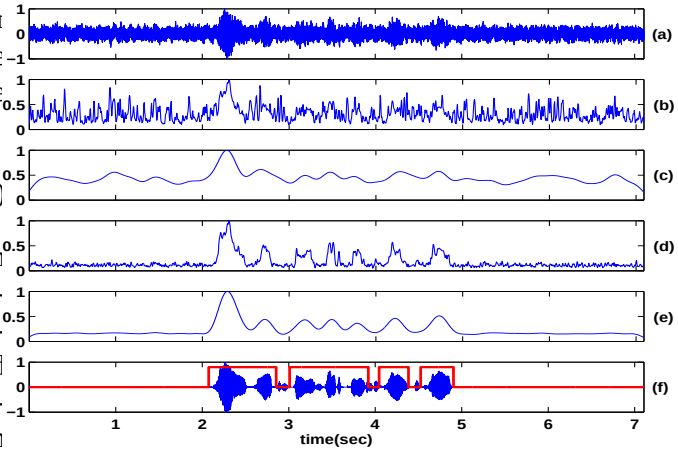

Figure 3: *The maximum contours of the DFT spectrum in pink noise at -5 dB; (a) Noisy signal, (b) Maximum amplitude in the magnitude spectrum, (c) Low-pass filtered signal of (b), (d) Maximum amplitude in the resonance 1 sub-band of magnitude spectrum, (e) Low-pass filtered signal of (d), (f) VAD from MSA-SB method (displayed on clean speech signal for reference)*

This motivated us to experiment with maximum contours in sub-bands that are specific to vocal tract resonances. The entire spectrum, is divided into three sub-bands, which were chosen to be 300-900Hz , 600-2800 Hz and 1400-3800 Hz corresponding to ranges of first three vocal tract resonances (Deng et al., 2006). The maximum in each sub-band of the spectrum is then computed. Fig. 2 (d) and 3 (d) show the maximum contour in resonance 1 sub-band corresponding to speech signal with white and pink noise at -5 dB. These maximum contours are then low-pass filtered (Figs. 2 (e) and 3 (e)). Thus, it can be seen that maximum contours in a sub-band specific to speech, is able to robustly discriminate speech and noise regions, as opposed to the full-band maximum contours. This is because maximum picked in sub-band 1 corresponds to vocal tract resonance in speech region and to an arbitrary maximum in noise regions. As transition of vocal tract is a continuum, the variation of maximum contour is smooth in speech regions and is otherwise in noise regions. And also in this sub-bands maximum of speech has higher amplitude than that of noise.

Experimental results show that maximum in sub-band 1 is sufficient for robust VAD. VAD decision is obtained by setting a threshold on the low-pass filtered version of maximum contour. Figs. 2 (f) and 3 (f) show the resulting VAD. One way of setting threshold is by picking a maximum in first 50 ms from low-pass filtered version of the noisy signal assuming that it is devoid of speech. This threshold automatically varies for different noises and SNRs. Though, it is the simplest way of selecting threshold, it might not be the appropriate way in all cases. Thus, for a more efficient thresholding operation, we used the combined decision of low-pass filtered versions of three bands. Mean subtraction and variance normalization is performed on low-pass filtered versions of three selected bands. The output is summed up and again mean subtraction and variance normalization is performed to get a final contour on which VAD decision is to be taken. The histogram for this final contour varies between -2 to 5. So, threshold is varied between -0.5 to 0.8 to decide upon a proper value for speech-nonspeech decision. ROC curves obtained are shown in fig. 4. We can observe that the same threshold that is independent of noise and SNR can be applied on final contour to get an appropriate VAD decision. This is due to combination of three sub-bands, followed by mean and variance normalization that is canceling the effect of noise level through out the signal. Sohn method (Sohn et al., 1999) for VAD provides an option to vary thresholds. False alarm rate (FAR) and correct detection rate (CDR) varies according to threshold. ROCs are plotted by taking FAR on x-axis and CDR on y-axis for various thresholds. ROCs of our method are compared with VAD using Sohn (Sohn et al., 1999) method as shown in fig. 4. It is observed that our method outperforms Sohn for all the tested noises at different dBs. After selecting an appropriate threshold from ROC, our method is compared with AMR-VAD2 (AMR, 1998) in the results section.

Table 2: *FAR and MR for various noises in different SNRs for MSA-SB and AMR2 methods*

White Noise	**5 dB**		**0 dB**		**-5 dB**		**-10 dB**	
	MR %	FAR %	MR %	FAR %	MR %	FAR %	MR %	FAR %
MSA-SB	12.71	1.98	**15.79**	1.80	**20.62**	1.59	**28.50**	1.34
AMR2	6.92	2.39	20.83	1.49	47.80	0.63	83.38	0.32

Pink Noise	**5 dB**		**0 dB**		**-5 dB**		**-10 dB**	
	MR %	FAR %	MR %	FAR %	MR %	FAR %	MR %	FAR %
MSA-SB	14.70	**1.85**	**19.24**	**1.61**	**26.57**	1.46	**39.50**	2.28
AMR2	5.84	2.80	21.32	1.83	50.83	0.72	83.24	0.48

HFchannel Noise	**5 dB**		**0 dB**		**-5 dB**		**-10 dB**	
	MR %	FAR %	MR %	FAR %	MR %	FAR %	MR %	FAR %
MSA-SB	14.92	**1.87**	19.58	**1.67**	**26.24**	**1.51**	**37.19**	2.13
AMR2	3.89	3.95	17.04	2.66	42.69	2.05	73.95	1.40

Factory1 Noise	**5 dB**		**0 dB**		**-5 dB**		**-10 dB**	
	MR %	FAR %	MR %	FAR %	MR %	FAR %	MR %	FAR %
MSA-SB	16.73	**3.43**	23.00	**6.63**	32.32	**11.99**	**39.39**	**22.61**
AMR2	2.87	37.54	10.13	36.25	25.84	37.70	48.47	37.36

Buccaneer1 Noise	**5 dB**		**0 dB**		**-5 dB**		**-10 dB**	
	MR %	FAR %	MR %	FAR %	MR %	FAR %	MR %	FAR %
MSA-SB	15.46	**1.83**	**19.83**	**1.66**	**26.30**	1.63	**41.46**	3.73
AMR2	7.17	3.29	24.36	2.05	56.17	1.18	87.66	0.84

Volvo Noise	**5 dB**		**0 dB**		**-5 dB**		**-10 dB**	
	MR %	FAR %	MR %	FAR %	MR %	FAR %	MR %	FAR %
MSA-SB	9.59	**2.21**	10.36	**2.13**	11.54	**2.04**	13.64	**1.93**
AMR2	0.54	5.57	0.50	5.39	0.51	5.20	0.93	5.14

Babble Noise	**5 dB**		**0 dB**		**-5 dB**		**-10 dB**	
	MR %	FAR %	MR %	FAR %	MR %	FAR %	MR %	FAR %
MSA-SB	22.15	**6.16**	28.51	**14.84**	40.53	**20.44**	54.75	**24.30**
AMR2	0.93	36.78	4.57	34.05	14.58	31.98	29.80	30.42

4.1 Results

The proposed algorithms are compared against the standard ETSI AMR-VAD2 (AMR, 1998). The FAR and MR of our methods along with the baseline techniques in various noisy conditions in four different SNRs are reported in Table 2. The corresponding HTER is plotted in Fig. 5. The lower HTER indicates better performance of the algorithm. We can observe from the bar graph that for most of the noise conditions, MSA-SB method outperforms (3^{rd} bar (light yellow) from the left in every noise) all other methods at low SNR levels. For volvo noise, we can see that MSA-SB method has lower FAR but higher MR compared

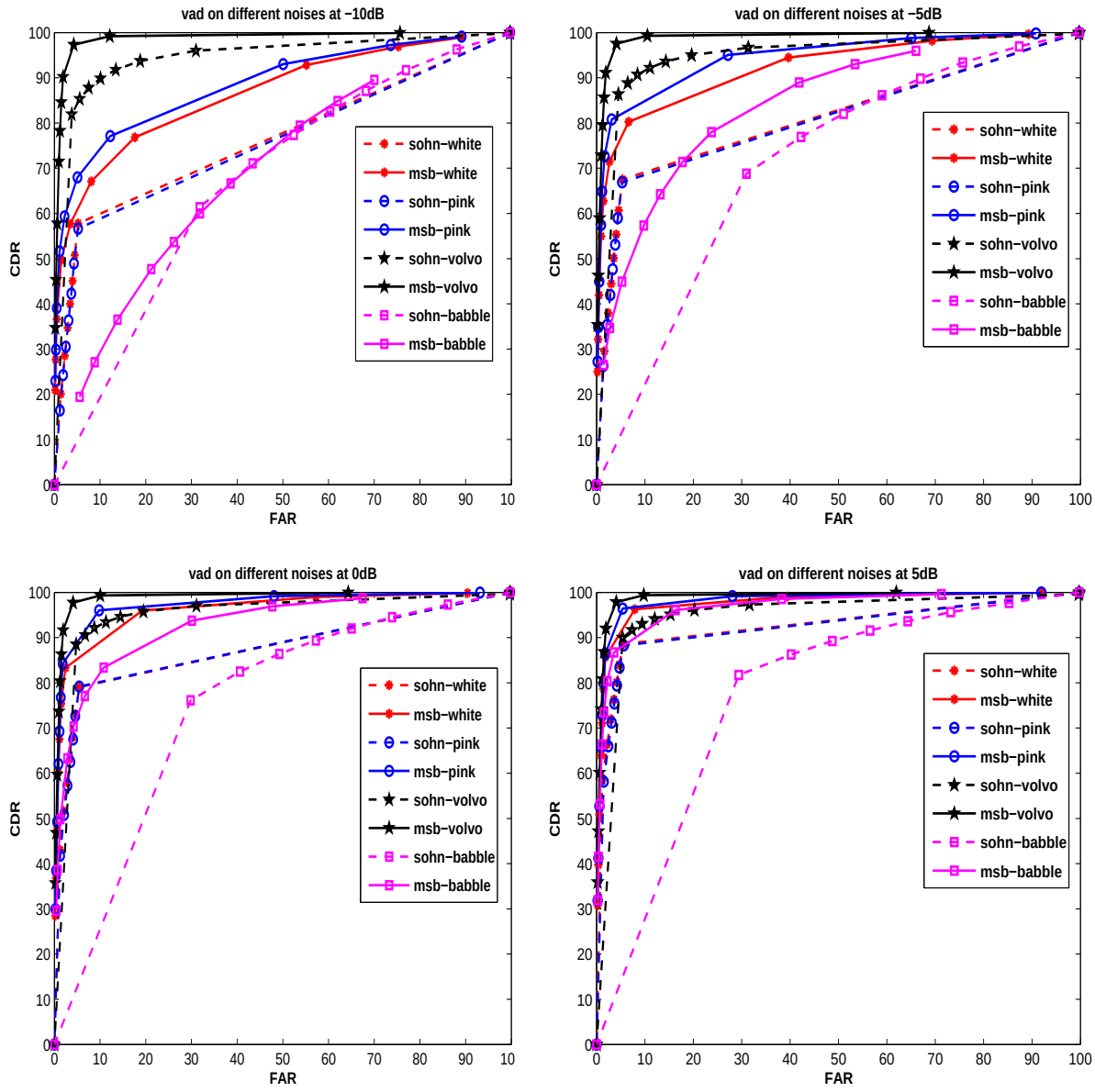

Figure 4: *ROC curves for different noises at -10, -5, 0 and 5 dB*

to AMR2. This is because of the threshold setting, some unvoiced and stop sounds might have been recognised as nonspeech in volvo. In babble noise, from the Table 2 one can observe that the FAR is consistently lower for MSA-SB than that of the AMR2 method. However in AMR2, MR is lower than all the methods for all SNRs in babble. This is attributed to the fact that our algorithms rely on speech specific features and babble being speech like noise, shows a drop in the performance. In summary, for most noises MSA-SB outperforms AMR2, while in some it performs comparable to it.

5 Conclusions and Future Work

In this paper, we investigated two methods for VAD in low SNR conditions. We experimented on seven noise conditions under four different SNRs. The time domain analysis revealed that lag of the autocorrelation function at peak (ACF-LAG) is more reliable than peak value (ACF-MAX) itself. The frequency domain MSA-SB method was found to be very robust even under very low SNR conditions and justifies our motivation for choosing sub-bands specific to vocal tract resonance ranges. The combination of sub-bands followed by mean and variance normalization has resulted in choosing a threshold independent of noise con-

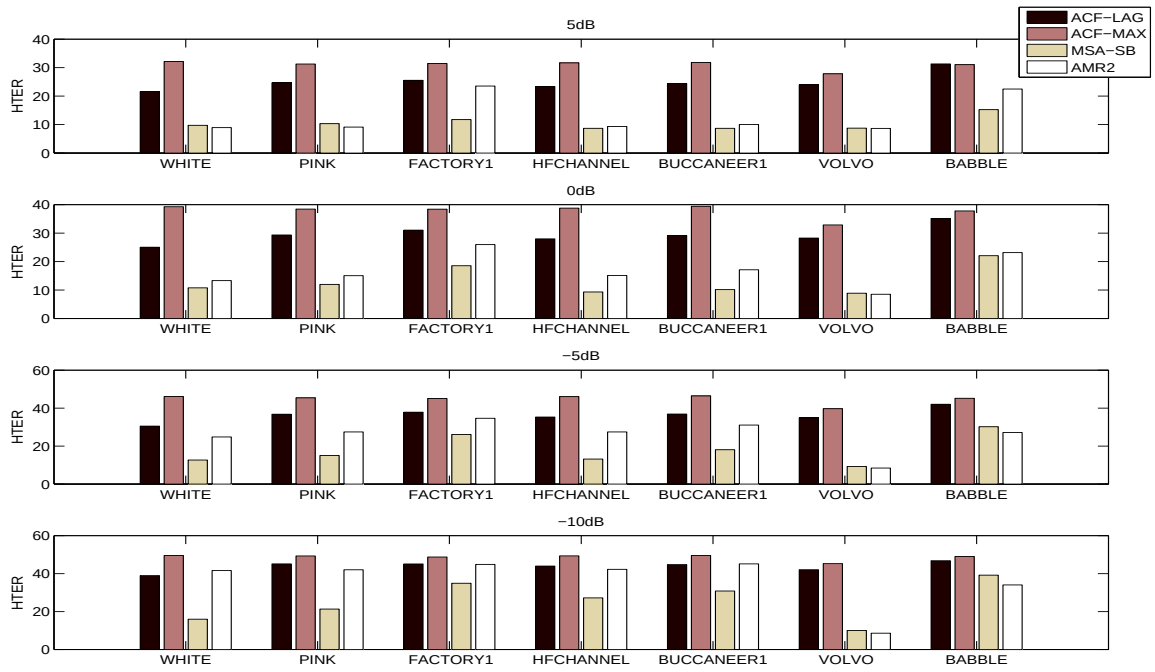

Figure 5: *% HTER performance of the proposed algorithms along with the baseline methods for each noise scenario at SNRs 5dB, 0 dB, -5dB, -10dB.*

ditions and levels. In future, we plan to do extensive evaluation of the technique on real world speech signals.

Acknowledgment

The authors would like to thank Sudarsana, Anand, Gautam, Vasanth, Bhargav and Santosh for their valuable feedback.

References

1998. Digital cellular telecommunications system (Phase 2+); Adaptive multi rate (AMR) speech; ANSI-C code for AMR speech codec.

F. Beritelli, S. Casale, and A. Cavallaero. 1998. A robust voice activity detector for wireless communications using soft computing. *IEEE Journal on Selected Areas in Communications*, 16(9):1818–1829, Dec.

Alain de Cheveign and Hideki Kawahara. 2002. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930.

Li Deng, A. Acero, and I. Bazzi. 2006. Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint. *IEEE Trans. Audio, Speech Lang. Process.*, 14(2):425–434, March.

Houman Ghaemmaghami, Brendan J Baker, Robert J Vogt, and Sridha Sridharan. 2010. Noise robust voice activity detection using features extracted from the time-domain autocorrelation function. *Proc., INTERSPEECH*.

P.K. Ghosh, A. Tsiartas, and S. Narayanan. 2011. Robust voice activity detection using long-term signal variability. *IEEE Trans. Audio, Speech Lang. Process.*, 19(3):600–613.

K. Itoh and M. Mizushima. 1997. Environmental noise reduction based on speech/non-speech identification for hearing aids. In *Proc., ICASSP*, volume 1, pages 419–422, Apr.

Lamia Karray and Arnaud Martin. 2003. Towards improving speech detection robustness for speech recognition in adverse conditions. *Speech Communication*, 40(3):261 – 276.

T. Kinnunen and P. Rajan. 2013. A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data. In *Proc., ICASSP*, pages 7229–7233.

J. Ramirez, J.C. Segura, C. Benirez, A. de la Torre, and A. Rubio. 2004. A new voice activity detector using subband order-statistics filters for robust speech recognition. In *Proc., ICASSP*, volume 1, pages 849–52, May.

J. Ramirez, J.C. Segura, C. Benitez, L. Garcia, and A. Rubio. 2005. Statistical voice activity detection

using a multiple observation likelihood ratio test. *IEEE Signal Process. Lett.*, 12(10):689–692, Oct.

S.O. Sadjadi and J.H.L. Hansen. 2013. Unsupervised speech activity detection using voicing measures and perceptual spectral flux. *IEEE Signal Process. Lett.*, 20(3):197–200, March.

Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. 1999. A statistical model-based voice activity detection. *IEEE Signal Process. Lett.*, 6(1):1–3, Jan.

Lee Ngee Tan, B.J. Borgstrom, and Abeer Alwan. 2010. Voice activity detection using harmonic frequency components in likelihood ratio test. In *Proc., ICASSP*, pages 4466–4469, March.

Peng Teng and Yunde Jia. 2013. Voice activity detection using convolutive non-negative sparse coding. In *Proc., ICASSP*, pages 7373–7377, May.

1993. DARPA-TIMIT. *Acoustic-Phonetic Continuous Speech Corpus*.

Andrew Varga and Herman J. M. Steeneken. 1993. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251.

Datao You, Jiqing Han, Guibin Zheng, and Tieran Zheng. 2012. Sparse power spectrum based robust voice activity detector. In *Proc., ICASSP*, pages 289–292, March.