Comparison of Voice Activity Detectors for Interview Speech in NIST Speaker Recognition Evaluation

Hon-Bill Yu and Man-Wai Mak

Department of Electronic and Information Engineering,
The Hong Kong Polytechnic University

Abstract

Interview speech has become an important part of the NIST Speaker Recognition Evaluations (SREs). Unlike telephone speech, interview speech has substantially lower signal-to-noise ratio, which necessitates robust voice activity detection (VAD). This paper highlights the characteristics of interview speech files in NIST SREs and discusses the difficulties in performing speech/nonspeech segmentation in these files. To overcome these difficulties, this paper proposes using speech enhancement techniques as a pre-processing step for enhancing the reliability of energy-based and statistical-model-based VADs. It was found that spectral subtraction can make better use of the background spectrum than the likelihood-ratio tests in statistical-model-based VADs. A decision strategy is also proposed to overcome the undesirable effects caused by impulsive signals and sinusoidal background signals. Results on NIST 2010 SRE show that the proposed VAD outperforms the statistical-model-based VAD, the ETSI-AMR speech coder, and the ASR transcripts provided by NIST SRE Workshop.

Index Terms: Voice activity detection, spectral subtraction, likelihood ratio tests, speaker verification, NIST SRE.

1. Introduction

Voice activity detection aims to detect speech and non-speech segments in audio signals. It is important to remove non-speech segments because they do not contain any speaker information. Early VAD methods extract features that appear only in speech segments and compare the quantified values of these features with a decision threshold (can be fixed or adaptive) for making speech/non-speech decisions [1]. The detection accuracy of these earlier methods, however, could degrade dramatically under adverse acoustic conditions. To address the robustness issues, methods that use the minima of noisy signals' power envelopes [1] and high-order cumulants of the LPC residual of short-term speech [2] have been proposed.

Advanced speech coders typically use more sophisticated methods in their VAD. For instance, in Option 2 of ETSI adaptive multi-rate (AMR) coder [3], VAD decisions depend on the energy of 16 frequency bands, background noise, channel SNR, frame SNR, and long-term SNR.

More recently, research has focused on statistical-model-based VAD where speech/non-speech decisions are based on the mean of the log-likelihood ratios between the observed signals and background noise in individual frequency bins [4]. To improve robustness under adverse acoustic environments, contextual information derived from multiple observations has been incorporated into the likelihood-ratio tests [5] and better soft-decision strategies and background noise estimation methods have been developed [6].

In recent NIST Speaker Recognition Evaluations (SREs), participating sites typically used energy features, the periodicity of speech frames, the power of noise-removed speech frames, and ASR transcripts provided by NIST in their VADs [7, 8, 9]. We have recently proposed an energy-based VAD that uses spectral subtraction as a preprocessor and showed promising results in NIST 2008 SRE [10]. The advantage of using spectral subtraction as a preprocessor is that it allows nonlinear filtering to be applied to the noisy signal, which effectively emphasizes the speech signal at high SNR regions and suppresses the background noise (to almost zero) in low SNR regions. This makes the subsequent energy-based VAD uncomplicated.

This paper extends our earlier work [10] in VAD to the interview speech and telephone speech recorded by different microphones in NIST 2010 SRE. Furthermore, the paper compares our proposed VAD with some state-of-the-art VADs, including the statistical-model-based VAD and the VAD in advanced speech coders.

2. Characteristics of Interview Speech in NIST SREs

The telephone speech files in early SREs have high signal-to-noise ratios (SNRs), making VAD a trivial task. In the interview speech of recent SREs, however, different microphone types were used for recording. For example, twelve microphones were used in NIST 2008 SRE, and in NIST 2010 SRE, the interviewees used different types of far-field microphones, such as lavaliere microphones, camcorders, and hanging microphones. These microphones lead to two types of speech files.

1. Files with extremely low SNR. Some of these files also contain low-energy speech superimposed on periodic background signals.
2. Files containing a number of spikes or impulsive signals caused by plosive sounds or puff of air produced by the speakers.

These characteristics are illustrated in Fig. 1 and Fig. 2.

3. Voice Activity Detection

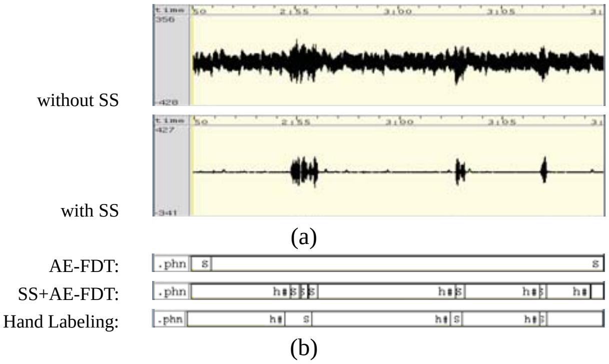
3.1. Statistical-Model Based VAD

Recent state-of-the-art VADs are based on likelihood ratio tests where the distributions of the frequency components of speech and noise are approximated by a statistical model (SM) [4]. The

This work was in part supported by RGC of Hong Kong (PolyU5264/09E) and the Hong Kong Polytechnic University (4-ZZ7W). We thank Helen Meng and Weiwu Jiang for providing their ETSI AMR coder.

VAD Method	Description
AE-FDT	Energy-based VAD with decision thresholds determined by the combination between average magnitude of background noise and signal peaks. The combination is controlled by a weighting factor (γ in Eq. 8).
ASR	Speech segments in the Automatic Speech Recognition transcripts provided by NIST [11].
AMR	VAD in ETSI Adaptive Multi-Rate coder (Option2) [3].
SM-ADT	Sohn's statistical-model-based VAD [4] incorporated with an adaptive threshold [12].
SS+SM-ADT	SM-ADT VAD with spectral subtraction as a pre-processing step.
SM-FDT	Sohn's statistical-model-based VAD incorporated with a fixed threshold, determined by Eq. 5.
SS+SM-FDT	SM-FDT VAD with spectral subtraction as a pre-processing step.
SS+AE-FDT	The proposed energy-based VAD with spectral subtraction as a pre-processing step.

Table 1: The voice activity detection (VAD) methods and their acronyms compared in this paper.



(a)

(b)

Figure 1: *(a) A short segment of an interview-utterance in NIST 2008 SRE without and with spectral subtraction (SS) as a pre-processor. (b) The segmentation results of an energy-based VAD without and with SS, and the segmentation done by listening tests (Hand Labeling). Labels* `s` *and* `h#` *represent speech and non-speech segments, respectively. See Table 1 for the meaning of acronyms.*

likelihood ratio for the kth frequency bin at frame n is

$$\Lambda_k(n) = \frac{1}{1+\xi_k(n)} \exp\left\{\frac{\gamma_k(n)\xi_k(n)}{1+\xi_k(n)}\right\} \qquad (1)$$

where $\gamma_k(n) = \frac{|Y_k(n)|^2}{\lambda_k(n)}$ is the *a posteriori* SNR between the noisy spectrum $Y_k(n)$ and the background variance $\lambda_k(n)$, and $\xi_k(n)$ is the decision directed *a priori* SNR

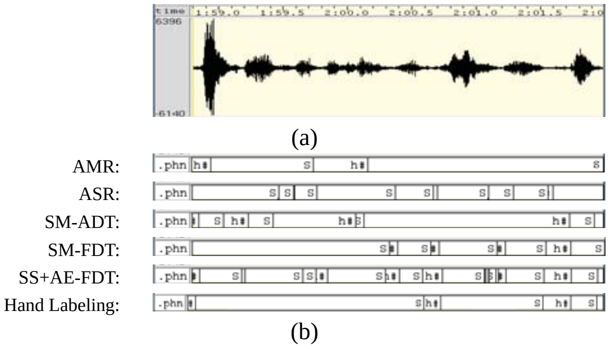
$$\xi_k(n) = \alpha \frac{\hat{X}_k^2(n-1)}{\lambda_k(n-1)} + (1-\alpha)P(\gamma_k(n)-1) \qquad (2)$$

where $\hat{X}_k(n-1)$ is the estimated spectrum of the previous frame obtained using the MMSE estimator. $P()$ is a function such that $P(x) = x$ if $x \geq 0$, and $P(x) = 0$ otherwise. In this work, $\lambda_k(n) = \lambda_k \forall n$.

To account for the correlation in consecutive speech frames, an HMM-based hangover scheme is adopted. Using this scheme, the decision rule can be written as [4]

$$\Gamma(n) = \frac{a_{01}+a_{11}\Gamma(n-1)}{a_{00}+a_{10}\Gamma(n-1)}\frac{P(H_0)}{P(H_1)}\Lambda(n) \underset{H_0}{\overset{H_1}{\gtrless}} \eta \qquad (3)$$

where η is a decision threshold; $a_{ij} \triangleq P(q(n) = H_j | q(n-1) = H_i), i \in \{0,1\}$, is the transition probability; and $\Lambda(n) = \left[\prod_{k=0}^{L-1} \Lambda_k(n)\right]^{\frac{1}{L}}$ is the geometric mean of the likelihood ratios of the individual frequency bins; and L is the frame size.



(a)

(b)

Figure 2: *(a) Waveform of a short speech segment with impulsive signals in NIST 2010 SRE. (b) Speech/non-speech decisions (*`s` *for speech and* `h#` *for silence) made by five different VAD abbreviated in Table 1 and listening tests (Hand Labeling).*

3.1.1. SM-VAD with adaptive threshold

The SM-based VAD uses the instantaneous SNR of individual frequency components in its decision process. This strategy makes better use of the background spectrum than the conventional energy-based VAD. Nevertheless, the VAD accuracy is still highly dependent on the decision threshold η, which is fixed across the whole utterance. One possible approach to alleviating this limitation is to make η adaptive so that it can track the fluctuation in the signal and noise power.

In this paper, SM-based VAD with adaptive decision thresholds (ADT) [12] is considered as a baseline VAD method. In this method, the statistical score $\Gamma(n)$ is compared with an adaptive decision threshold $\eta(n)$, which is updated in a frame-by-frame basis. More specifically,

$$\eta(n+1) = \mu\eta(n) + (1-\mu)\left(\overline{\Lambda}+\beta(n)\sigma_\Lambda\right) \qquad (4)$$

where $0 < \mu < 1$ is a forgetting factor and $\overline{\Lambda}$ and σ_Λ are the mean and standard deviation of $\Gamma(n)$ in Eq. 3 during non-speech region (determined by an energy-ranking approach to be discussed next). $\beta(n)$ is a weighting factor that will be increased if $\Gamma(n) > 1.25\eta(n)$; otherwise $\beta(n)$ will be decreased.

3.1.2. SM-VAD with fixed threshold

In this approach, the SM scores $\Gamma(n)$ of the entire utterance are ranked in descending order as shown in Fig. 3. Then, a fixed percentage of scores in the lower and upper ends of the ranked list are selected. The VAD's fixed decision threshold (FDT) is a linear combination of the score mean of the lower end (μ_b) and

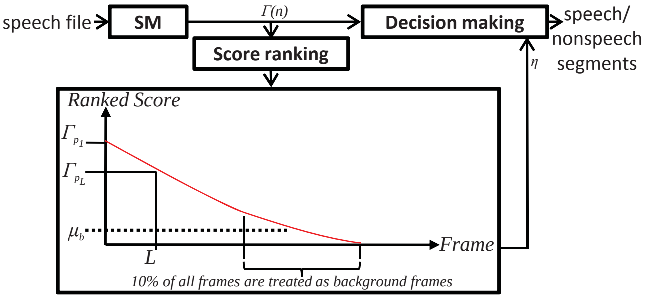


Figure 3: *The structure of SM-VAD incorporated with a fixed threshold (SM-FDT VAD) for NIST SREs.*

the minimum score in the upper end as follows:

$$\eta = \gamma\mu_b + (1-\gamma)\min\{\Gamma_{p_1},\ldots,\Gamma_{p_L}\}, \tag{5}$$

where $0 \ll \gamma < 1$ is a weighting factor and $\{\Gamma_{p_1},\ldots,\Gamma_{p_L}\}$ are scores of L frames with the largest value.

The above procedure raises the issue of determining an appropriate percentage for the lower and upper ends of the ranked score list. These percentages can be founded by inspecting several interview speech files in NIST 2005–2008 SREs. By examining some of these files, we found that it is fairly safe to consider 10% of a speech file contain background frames and 5% of the file contain signal peaks.

3.2. Noise-Reduced VAD

In this method, spectral subtraction (SS) with a large over-subtraction factor is used to remove the background noise before passing the enhanced speech to an energy-based VAD. To obtain the enhanced speech $\hat{x}(t)$ from the noisy speech $y(t)$ at frame m, we implemented the spectral subtraction [13] of the form

$$\hat{X}(\omega,m)=\begin{cases}\left[|Y(\omega,m)|-\alpha_m|\hat{B}(\omega)|\right]e^{j\varphi_y(\omega,m)} & \\ \qquad\qquad \text{if } |Y(\omega,m)|>(\alpha_m+\beta_m)|\hat{B}(\omega)| & \\ \beta_m|\hat{B}(\omega)|e^{j\varphi_y(\omega,m)} & \\ \qquad\qquad \text{otherwise,} & \end{cases} \tag{6}$$

where $\varphi_y(\omega,m)$ is the phase of $Y(\omega,m)$, $\hat{B}(\omega)$ is the average spectrum of some non-speech regions, $\alpha_m \geq 1$ is an over-subtraction factor for removing background noise, and $0 < \beta_m \ll 1$ is a spectral floor factor ensuring that the recovered spectra never fall below a preset minimum. The value of α_m and β_m can be computed as

$$\begin{aligned}\alpha_m &= -\frac{1}{2}\xi_m + c \qquad (\alpha_{\text{min}} \leq \alpha_m \leq \alpha_{\text{max}})\\ \beta_m &= \begin{cases}\beta_{\text{min}} & \text{if } \xi_m < 1\\ \beta_{\text{max}} & \text{otherwise}\end{cases}\end{aligned} \tag{7}$$

where $\xi_m = |Y(\omega,m)|^2/|\hat{B}(\omega)|^2$ is the *a posteriori* SNR, c is a constant ($= 4.5$ in this work) which ensures α_m will attain α_{max} when ξ_m is closed to zero, α_{min}, α_{max}, β_{min}, and β_{max} constrain the allowable range of the over-subtraction factor and the noise floor. In this work, we set $\alpha_{\text{max}} = 4$, $\alpha_{\text{min}} = 0.5$, $\beta_{\text{max}} = 0.05$, and $\beta_{\text{min}} = 0.01$.

The presence of spikes in some files needs to be taken care of when determining the VAD decision threshold. In particular, these spikes lead to overestimation of the decision threshold if it is based on the background amplitude and the maximum amplitude. To address this problem, we have developed a strategy that is similar to the FDT described in Section 3.1.2; however, instead of statistical scores, this strategy considers signal amplitudes. More specifically, the decision threshold is a linear combination of the minimum of the signal peaks and the mean of background amplitude:

$$\theta = \gamma\mu_b + (1-\gamma)\min\{a_{p_1},\ldots,a_{p_L}\}, \tag{8}$$

where $\{a_{p_1},\ldots,a_{p_L}\}$ are the amplitudes of L frames with the largest value. In this work, L was set to 1% of the total number of frames in the speech file. By comparing the amplitude of each frame in the file with the threshold, those frames with amplitude larger than the threshold are considered as speech frames.

4. Experiments

NIST 2005–2010 SREs were used in the experiments. NIST05–08 SREs were used as development data, and NIST10 was used for performance evaluations.[1] Only male speakers in these corpora were used. For each utterance in NIST10, eight VADs (see Table 1) were applied to remove the silence segments.

The weighting factor γ in Eq. 8 was set to 0.95 and 0.96 for AE-FDT and SS+AE-FDT, respectively. For SM-FDT and SS+SM-FDT VADs, γ in Eq. 5 was set to 0.993.

We extracted 12 MFCCs and their first derivatives from the speech regions of the utterances to create 24-dim acoustic vectors. Cepstral mean normalization was applied to the MFCCs, followed by feature warping.

The target-speakers were modeled by GMM-SVM [15]. In the modeling process, a gender-dependent universal background model (512-center) was created by using the interview utterances of NIST05–06. MAP adaptation, with relevance factor set to 16, was then performed for each of the target-speakers to create target-dependent GMMs. The same MAP adaptation was also applied to 300 background speakers (also from NIST05–06) to create 300 impostor GMMs.

The utterances of 144 male speakers from NIST05–08 were used for estimating the gender-dependent NAP matrices to reduce channel effects (NAP corank was set to 128). For the T-norm speaker models, 300 male utterances from NIST05 were used. The same set of background speakers used for creating the target-speaker SVMs were used for creating the T-norm SVMs.

5. Results and Discussions

Table 2 shows the equal error rate (EER) and minimum normalized decision cost (minNDCF) achieved by eight VAD methods. The results strongly suggest that preprocessing the noisy sound files by spectral subtraction (SS) is a promising idea. Specifically, spectral subtraction can reduce the overall EERs of AE-FDT and SM-FDT by 56% and 5%, respectively.

Comparing the results of AE-FDT and SS+AE-FDT reveals that SS has significant contribution to the conventional energy-based VAD. However, the performance of SS+SM-FDT is better than SM-FDT by a small margin only. This suggests that SS is not detrimental to the statistical-model-based VAD. The reason is that in SM-based VADs, the background spectrum has already been taken into account in the scoring function.

[1] Hereafter, all NIST SREs are abbreviated as NISTXX, where XX stands for the year of evaluation.

VAD Method	Speaker Modeling	EER (%)						Minimum Normalized DCF					
		CC1	CC2	CC4	CC7	CC9	Overall	CC1	CC2	CC4	CC7	CC9	Overall
AE-FDT	SVM	6.57	11.72	7.23	12.28	7.44	10.30	0.84	0.99	0.96	0.84	0.97	0.97
ASR	SVM	5.15	8.58	7.74	12.81	5.74	8.88	0.78	0.85	0.74	0.88	0.77	0.90
AMR	SVM	4.44	8.05	9.44	12.85	5.98	9.61	0.81	0.85	0.80	0.77	0.55	0.90
SM-ADT	SVM	4.21	6.61	6.84	11.56	5.63	7.16	0.74	0.77	0.76	0.79	0.49	0.86
SS+SM-ADT	SVM	4.24	7.88	6.93	9.27	5.10	8.59	0.60	0.78	0.78	0.79	0.62	0.96
SM-FDT	SVM	3.23	4.68	4.49	9.48	3.06	5.03	0.66	0.68	0.70	0.65	0.38	0.77
SS+SM-FDT	SVM	2.83	4.45	4.04	7.58	2.56	4.80	0.62	0.61	0.70	**0.59**	0.42	0.76
SS+AE-FDT	SVM	2.82	**4.44**	**3.51**	6.70	**2.37**	**4.55**	0.70	0.58	**0.62**	0.64	**0.17**	**0.72**
	JFA	**2.72**	4.75	3.90	**6.14**	4.27	5.56	**0.46**	**0.53**	0.67	0.68	0.22	0.75

Table 2: Performance on NIST 2010 SRE under Common Conditions (CC) 1, 2, 4, 7 & 9. **AE-FDT**: energy-based VAD without noise removal; **ASR**: VAD segmentation from NIST provided ASR transcripts; **AMR**: VAD in ETSI-AMR coder; **SM-ADT** & **SM-FDT**: Sohn's VAD [4] incorporated with adaptive threshold [12] and the proposed fixed threshold (Eq. 5); **SS+SM-ADT** & **SS+SM-FDT**: SM-ADT and SM-FDT with spectral subtraction; **SS+AE-FDT**: the proposed spectral-subtraction VAD; **SS+AE-FDT (with JFA)**: the proposed VAD was applied to a Joint Factor Analysis-based speaker verification system [14].

Note that SS+AE-FDT and SM-FDT use the background spectrum in different manners. For the former, the background spectrum is used for spectral subtraction, whereas for the latter it is used for computing the likelihood ratio scores. This difference enables us to make better use of the background spectrum in SS+AE-FDT. Specifically, to remove as much background noise as possible, we may apply a large upper-limit for the over-subtraction factor (α_{max}) and a small lower-limit for the noise floor (α_{min}).[2] The over-subtraction factor α_m is a linear function of the a posteriori SNR for certain range of SNR and is bounded by the lower- and upper-limit when the SNR is beyond this range. As a result, more background noise will be removed during low SNR region whereas more speech content will be retained during high SNR region. The SM-FDT, on the other hand, does not have such property because the background spectrum is assumed constant for both low and high SNR.

The results show that using the ASR transcripts provided by NIST SRE Workshop as VAD leads to poor speaker verification performance, suggesting that the ASR system does not produce accurate speech/non-speech segmentations. The VAD in ETSI-AMR coder also performs poorly. This is mainly caused by the overestimation of both the speech onset and offset regions.

6. Conclusions

To extract speech segments from the interview-speech files in NIST SREs, a voice activity detector is specially designed and evaluated under the NIST 2010 SRE protocol. This study finds that: (1) noise removal is of primary importance for VAD under extremely low SNR, (2) a reliable threshold strategy is required for spiky (impulsive signal) speech, and (3) our proposed spectral subtraction VAD outperforms the ASR transcripts provided by NIST, the VAD in the advanced speech coder (ETSI-AMR, Option2) and the state-of-the-art statistical-model-based VAD in speaker verification.

7. References

[1] 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. 2, pp. 109–118, Feb 2002.

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

[3] ETSI, *Voice activity detector VAD for adaptive multi-rate (AMR) speech traffic channels, ETSI EN 301 708 v7.1.1*, 1999.

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

[5] J. Ramirez, J. C. Segura, J. M. Gorriz, and L. Garcia, "Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition," *IEEE Trans. Audio Speech Language Processing*, vol. 15, no. 8, pp. 2177–2189, 2007.

[6] I. Mica, H. Atassi, J. Prinosil, and P. Novak, "Voice activity detection under the highly fluctuant recording conditions of call centres," in *Proceedings of ECS'10/ECCTD'10/ECCOM'10/ECCS'10*, 2010, pp. 334–336.

[7] V. Hautamaki, M. Tuononen, T. Niemi-Laitinen, and P. Franti, "Improving speaker verification by periodicity based voice activity detection," in *Proc. 12th Int. Conf. Speech and Computer (SPECOM'2007)*, Moscow, October 2007, vol. 2, pp. 645–650.

[8] H. Sun, B. Ma, and H. Li, "An Efficient Feature Selection Method for Speaker Recognition," in *ISCSLP'08*, 2008, pp. 1–4.

[9] E. Dalmasso, F. Castaldo, P. Laface, D. Colibro, and C. Vair, "Loquendo - politecnico di torino's 2008 NIST speaker recognition evaluation system," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009*, Taipei, April 2009, pp. 4213–4216.

[10] M. W. Mak and H. B. Yu, "Robust voice activity detection for interview speech in NIST speaker recognition evaluation," in *Proc. APSIPA ASC 2010*, Singapore, 2010.

[11] A. Martin and C. Greenberg, Eds., *NIST SRE10 workshop*, Brno, Czech Republic, June 2010.

[12] S. S. Ahn and Y. C. Lee, "An improved statistical model-based VAD algorithm with an adaptive threshold," *Journal of the Chinese Institute of Engineers*, vol. 29, no. 5, pp. 783–789, 2006.

[13] J. R. Deller Jr, J. G. Proakis, and J. H. L. Hansen, *Discrete-time Processing of Speech Signals*, Macmillan Pub. Company, 1993.

[14] W. Rao and M. W. Mak, "Addressing the data-imbalance problem in kernel-based speaker verification via utterance partitioning and speaker comparison," in *Interspeech'11*, Florence, August 2011.

[15] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, pp. 308–311, 2006.

[2] Note that musical noise is not a concern because the denoised speech is only used for VAD, not for speaker recognition.