Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 28 (2014) 295–313

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

A study of voice activity detection techniques for NIST speaker recognition evaluations

Man-Wai Mak *, Hon-Bill Yu

Centre for Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China

Received 18 May 2012; received in revised form 4 June 2013; accepted 21 July 2013
Available online 31 July 2013

Abstract

Since 2008, interview-style speech has become an important part of the NIST speaker recognition evaluations (SREs). Unlike telephone speech, interview speech has lower signal-to-noise ratio, which necessitates robust voice activity detectors (VADs). This paper highlights the characteristics of interview speech files in NIST SREs and discusses the difficulties in performing speech/non-speech 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. A decision strategy is also proposed to overcome the undesirable effects caused by impulsive signals and sinusoidal background signals. The proposed VAD is compared with the ASR transcripts provided by NIST, VAD in the ETSI-AMR Option 2 coder, satistical-model (SM) based VAD, and Gaussian mixture model (GMM) based VAD. Experimental results based on the NIST 2010 SRE dataset suggest that the proposed VAD outperforms these conventional ones whenever interview-style speech is involved. This study also demonstrates that (1) noise reduction is vital for energy-based VAD under low SNR; (2) the ASR transcripts and ETSI-AMR speech coder do not produce accurate speech and non-speech segmentations; and (3) spectral subtraction makes better use of background spectra than the likelihood-ratio tests in the SM-based VAD. The segmentation files produced by the proposed VAD can be found in http://bioinfo.eie.polyu.edu.hk/ssvad.

Keywords: Speaker verification; Voice activity detection; NIST SRE; Statistical model based VAD; Spectral subtraction

1. Introduction

NIST speaker recognition evaluations (SREs) have been focusing on text-independent speaker verification over telephone channels since 1996. In recent years, NIST introduces interview-style speech into the evaluations. For example, the speech files in NIST 2008 SRE contain conversation segments of approximately 5 min of telephone speech and 3 min of interview speech, and the speech files in NIST 2010 SRE contain interview recordings with duration ranging from 3 to 15 min. In each speech file, about half of the conversation contains speech, and the remaining part contains pauses or silence intervals. The inclusion of non-speech intervals in the speech files necessitates voice activity detection (VAD) because these intervals do not contain any speaker information. In particular, VAD can be used to identify speech segments prior to the feature extraction process.

* Corresponding author. Tel.: +852 2766 6257.

0885-2308/$ – see front matter
http://dx.doi.org/10.1016/j.csl.2013.07.003

Speech/non-speech detection can be formulated as a statistical hypothesis problem aimed at determining to which class a given speech segment belongs. However, a high level of background noise can cause numerous detection errors, because the noise partly or completely masks the speech signal (Ramirez et al., 2007). A robust decision rule that works under noisy conditions is therefore essential. Most of the existing VAD algorithms are effective under clean acoustic environments, but they could fail badly under adverse acoustic conditions (Beritelli et al., 2002).

Traditionally, VAD uses periodicity measure (Tucker, 1992), zero-crossing rate (Benyassine et al., 1997), pitch (Chengalvarayan, 1999), energy (Woo et al., 2000), spectrum analysis (Marzinzik and Kollmeier, 2002), higher order statistics in the LPC residual domain (Nemer et al., 2001), or combinations of different features (Tanyer and Ozer, 2000). More sophisticated VAD techniques have been proposed for real-time speech transmission on the Internet (Sangwan et al., 2002) and mobile communication services (Freeman et al., 1989). In particular, the adaptive multi-rate (AMR) codec option II (AMR2) (ETSI, 2001) uses a decision logic based on the energy of 16 frequency bands, background noise, channel SNR, frame SNR, and long-term SNR (Cornu et al., 2003). The VAD of this codec takes advantage of speech encoder parameters and is more robust against environmental noise than its earlier version (AMR1) and G.729 (Torre et al., 2006). Moreover, the VAD decision threshold can be adapted dynamically according to the acoustic environment, allowing on-line speech/non-speech detection under non-stationary acoustic environments.

More recently, research has focused on statistical-model-based VAD where individual frequency bins of speech are assumed to follow a parametric density function (Sohn et al., 1999). In this approach, VAD decisions are based on likelihood ratio tests (LRTs) where the geometric mean of the log-likelihood ratios of individual frequency bins are estimated from observed speech signals. The statistical model can be Gaussian (Sohn et al., 1999) or generalized Gaussian (Góriz et al., 2010). However, it has been recently found that Laplacian and Gamma models are more appropriate for handling a wide variety of noise conditions (Chang et al., 2006). Using an online version of the Kolmogorov–Smirnov test, the type of models can be selected adaptively for different noise types and SNRs (Chang et al., 2006). To improve the robustness of VAD under adverse acoustic environment, contextual information derived from multiple observations has been incorporated into the LRT (MO-LRT) (Ramirez et al., 2007). Gaussian mixture models have been applied to model the static harmonic-structure information and the long-term temporal information of speech. VAD decisions are then based on the log-likelihood ratios computed from the clean and noise GMMs (Torre et al., 2006; Fukuda et al., 2010). In Sun et al. (2009), Wiener filtering is applied to remove noise before extracting acoustic features for training the speech and non-speech GMMs.

Characteristics of speech and non-speech signals have also been modeled by hidden Markov models (HMMs). For example, in Varela et al. (2011), a decision-tree algorithm that combines the scores of HMM-based speech/non-speech models and speech pulse information was used for rejecting far-field speech in speech recognition systems. Both Fukuda et al. (2010) and Varela et al. (2011), Vlaj et al. (2012) use statistical models to characterize speech and non-speech signals, with some decision logics governing the switching between speech and non-speech states. The difference being that in the GMM-VAD of Fukuda et al. (2010), state duration is governed by the number of speech frames (as detected by the GMMs) in a fixed-length buffer, and that in the GMM-VAD of Vlaj et al. (2012) state duration is governed by a hangover and handbefore scheme which detects the consonants occurred at the beginning, middle and the end of words; whereas in the HMM-VAD of Varela et al. (2011), the state duration is controlled by the state-transition probabilities of the HMMs and speech pulse information. Note that both GMM- and HMM-based VADs require ground-truth speech/non-speech segments for training the statistical models. Unfortunately, these labeled segments are not available in NIST SREs.

The VAD problem has also been formulated as an edge-detection problem. For example, in Li et al. (2002), two optimal 1D filters with responses invariant to various background noise levels are designed to detect the beginning edges and ending edges of the energy profile of speech signals. To detect the beginning edges, the filter has positive response to a beginning edge, negative response to an ending edge, and near-zero response to silence. The filter for detecting the ending edge has the opposite characteristics and has more time points. The filters are operated as a moving-average filter on the energy envelope and their outputs are compared with dynamic decision thresholds estimated from a 2-mode Gaussian mixture model.

In recent NIST SREs, several sites provided the details of their VAD in the system descriptions. Typically, these systems use energy-based methods that estimate a file-dependent decision threshold according to the maximum energy level of the file (Kinnunen et al., 2009). Some sites used the periodicity of speech frames or the power of noise-removed speech frames to make speech/non-speech decisions (Hautamaki et al., 2007; Sun et al., 2008; Mak and Yu, 2010; Yu

and Mak, 2011). An alternative approach is to use the ASR transcripts supplied by NIST to remove the non-speech segments (Dalmasso et al., 2009).

In this paper, we propose a VAD that is specifically designed for NIST SREs. Special attention has been paid to address the low SNR, impulsive noise, and cross talks in the interview-style speech files. The main idea is to apply speech enhancement as a pre-processing step to boost the SNR of the speech segments, which facilitates the subsequence speech/non-speech decisions either by log-likelihood ratio tests or comparing with energy-based thresholds. While this strategy has been adopted in the past, e.g., Sun et al. (2009), Ramirez et al. (2004), and Marciniak et al. (2008), our proposed VAD has some important differences. For example, the VAD in Sun et al. (2009) requires the training of speech and non-speech GMMs, whereas ours does not require training. This requirement is a burden for situations like NIST SREs because labeled speech segments are not available. The Wiener filtering in Sun et al. (2009) and Ramirez et al. (2004) and the wavelet denoising in Marciniak et al. (2008) also need to strike a good balance between spectral distortion and the degree of noise removal because decisions of these VADs are based on the spectral features of the noise-reduced speech. Our VAD, on the other hand, does not use the spectral features for VAD decisions. Therefore, it can leverage the over-subtraction to boost the SNR for better discrimination between speech and non-speech. To the best of our knowledge, our study provides the first comprehensive comparison between different VADs for NIST SREs. Results based on NIST 2010 SRE suggest that the proposed VAD outperforms the VAD in AMR2, the transcriptions provided by NIST, and statistical model-based VAD.

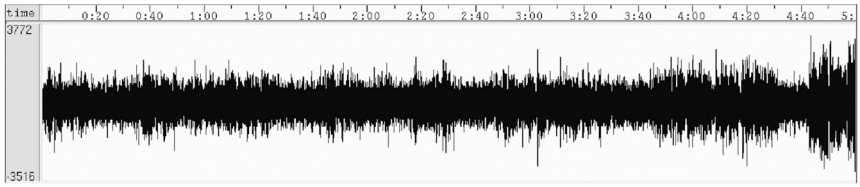
In Section 2, we highlight the characteristics of the interview speech files in NIST SRE and explain why conventional VAD techniques will encounter difficulty in detecting speech in these files. Then, in Section 3, we outline two state-of-the-art statistical VADs and explain how they can be applied to NIST SREs. Section 4 proposes using speech enhancement techniques as a pre-processing step for improving the statistic model based VAD and energy-based VAD. Experimental evaluations comparing different types of VADs under NIST 2010 SRE are then presented in Section 5.

2. Characteristics of interview speech in NIST SREs

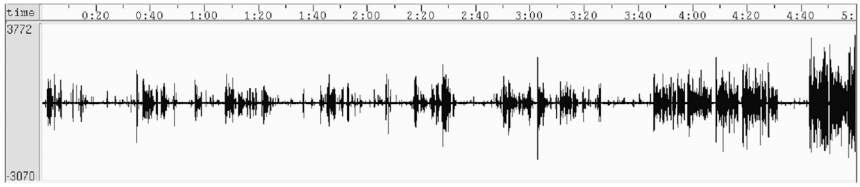
In early NIST SREs, researchers seldom pay attention to VAD. This is because the telephone speech files in early SREs have high signal-to-noise ratios (SNRs), making VAD a trivial task. The high SNR in telephone speech is resulted from the close proximity between speaker's mouth and the handset. In interview speech, however, different microphone types were used for recording. For example, 12 microphones were used in NIST 2008 SRE,[1] and in NIST 2010 SRE, the interviewees used different types of far-field microphones, such as lavaliere microphones, camcorders, and hanging microphones (Martin and Greenberg, 2010). These microphones lead to files with the following characteristics:

1 *Low SNR*. Depending on the microphone types, some of the interview speech segments have low SNR, causing problems in conventional VAD. Fig. 1(a) shows the waveform of an interview speech file (`ftvhv.sph`) in NIST 2008 SRE, and Fig. 1(c) highlights a short segment of the same file. The NIST STNR Tool[2] indicates that the SNR of this file is 5 dB. Although this level of SNR is not very low, it already causes numerous errors in an energy-based VAD, as indicated by "AE-VAD" in the lower panel of Fig. 1(c). Fig. 2 shows the histograms of SNR of interview speech files in NIST 2008 and 2010 SRE. While the mean SNRs of these two databases are high (22 dB and 21 dB, respectively), about 2% of the files have SNR less than 5 dB, i.e., about 2% of the files have situation similar to Fig. 1. The VAD errors in these files will have detrimental effect on speaker verification performance, which will be demonstrated in Section 5.
2 *Impulsive*. Some of the files in NIST 2010 SRE contain a large number of spikes that seriously mask the amplitude of speech segments, as illustrated in Fig. 3.
3 *Low-energy speech superimposed on periodic background signals*. Some files contain low-energy speech superimposed on periodic background noise, as exemplified in Fig. 4.
4 *Cross talk*. Each interview speech file in NIST 2010 SRE contains two channels, one recording the speech of an interviewee and the other the speech of an interviewer. As far-field microphones were used for recording interviewee's

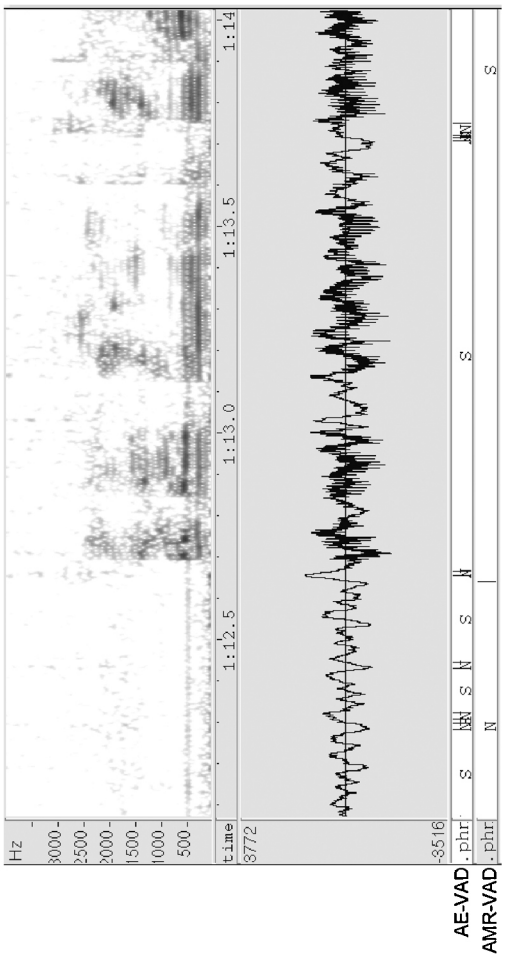
[1] Some of these microphones are of the same models, but they were placed at different positions with respect to the speakers.
[2] This tool is part of the Speech File Manipulation Software(SPHERE) Package Version 2.7, available from http://www.nist.gov/itl/iad/mig/tools.cfm.



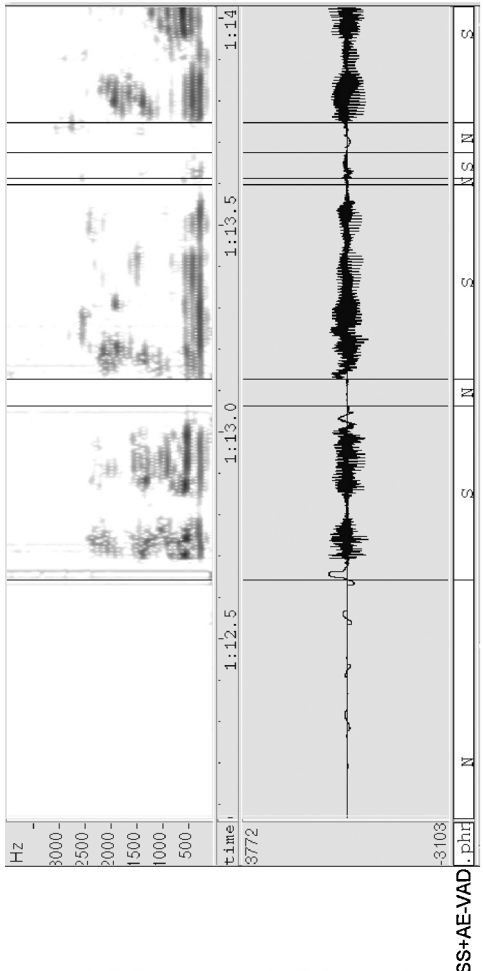
(a) The whole speech file (without denoising)



(b) The whole speech file (with denoising)



(c) A short segment (without denoising)



(d) A short segment (with denoising)

Fig. 1. Waveform, spectrogram, and speech/non-speech decision of an energy-based VAD and the ETSI-AMR coder on an interview speech file. (a and c) Without denoising and (b and d) with denoising. The VAD decisions (S for speech and N for non-speech) are shown in the bottom panels of (c and d). See Table 1 for the abbreviations of the VADs.

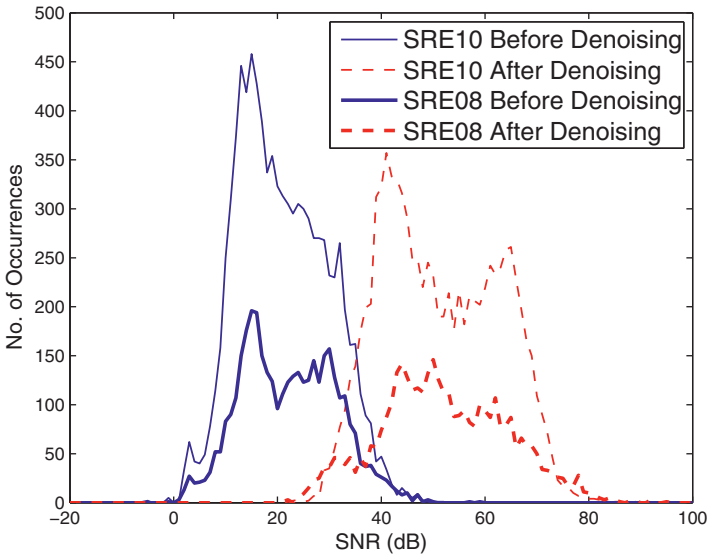


Fig. 2. Histograms of SNR of interview speech files in NIST 2008 and 2010 SREs before and after spectral subtraction. The SNRs were measured by the NIST STNR Tool.

speech, a low-energy crosstalk signal appears in the interviewee's channel when the interviewer is talking, causing the VAD mistakenly considers the crosstalk as belonging to the interviewee. This situation is exemplified in Fig. 5(a) in which the microphone of the interviewee's channel picks up the speech of the interviewer in Interval A.

As shown in these figures, conventional energy-based VAD fails to detect the speech segments under such conditions.

3. Statistical voice activity detection

This section highlights the merit of the statistical model based VAD (Sohn et al., 1999) and GMM-based VAD (Fukuda et al., 2010) and explains how they can be applied to detect the speech segments of NIST SRE speech files. The section focuses on the decision logic and threshold determine methods that are specifically designed for the SREs.

3.1. Statistical model (SM) based VAD

In SM-based VAD (Sohn et al., 1999), speech/non-speech segmentation is formulated as a hypothesis testing problem:

$$\begin{aligned} H_0 &: \text{speech absent} : Y(m) = B(m) \\ H_1 &: \text{speech present} : Y(m) = X(m) + B(m) \end{aligned} \tag{1}$$

where $Y(m)$, $X(m)$, and $B(m)$ represent the DFT of noisy speech, clean speech, and background noise at frame m, respectively. The complex DFT coefficients in $Y(m)$, $X(m)$, and $B(m)$ are assumed to be independent and normally distributed. For each frame m, a VAD score $\Gamma(m)$ is computed based on the VAD score of the previous frame and the likelihood ratio $\Lambda(m)$ at the current frame:

$$\Gamma(m) = \frac{P(H_0)}{P(H_1)} \left[\frac{a_{01} + a_{11}\Gamma(m-1)}{a_{00} + a_{10}\Gamma(m-1)} \right] \Lambda(m) \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{2}$$

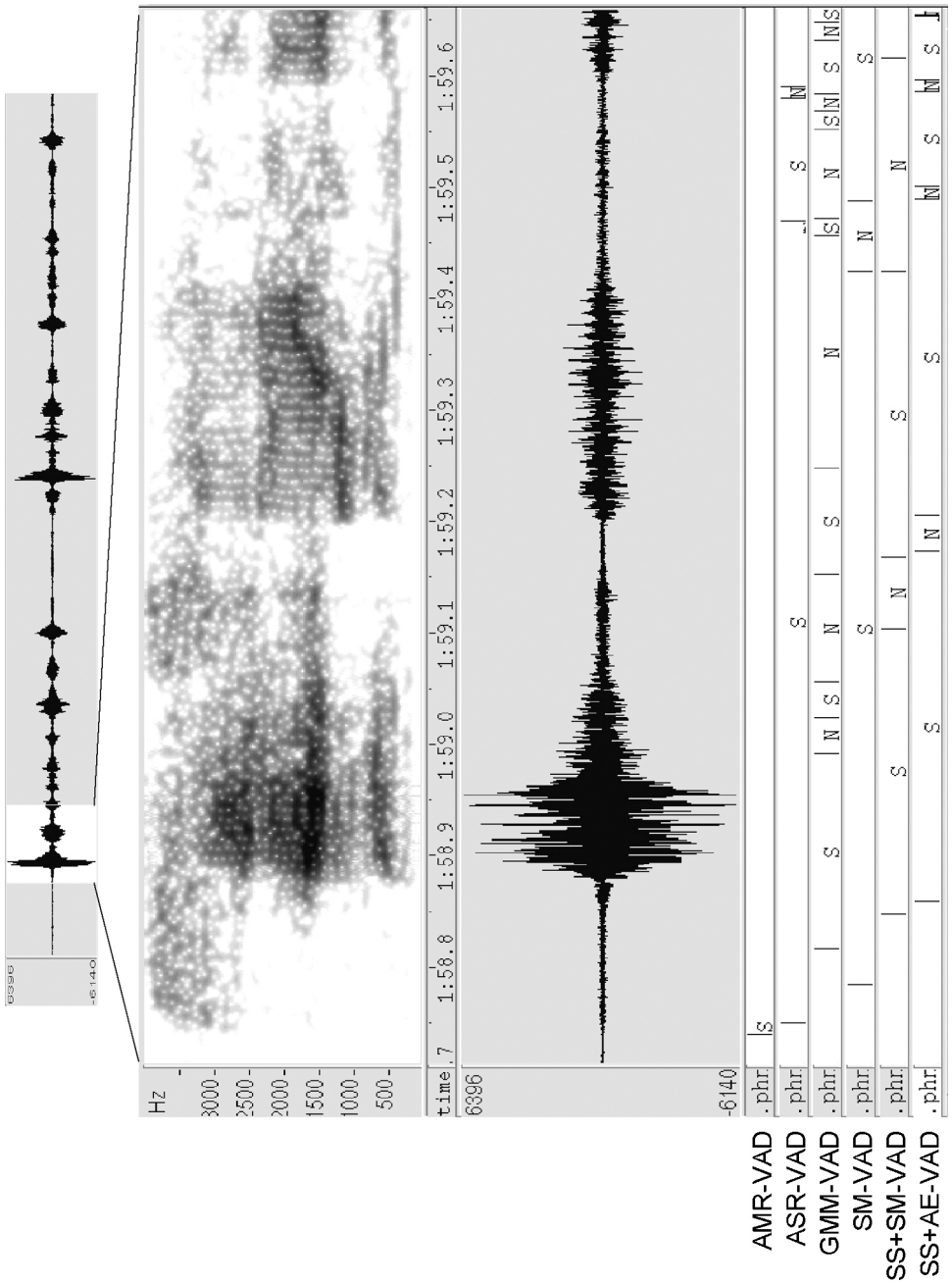


Fig. 3. A short segment of low-energy interview speech in NIST 2010 SRE with high-energy spikes. The bottom panel shows the speech (S) and non-speech (N) decisions made by six VADs detailed in Table 1.

where $a_{ij} \triangleq \Pr(q(m) = H_j | q(m-1) = H_i)$ are state-transition probability and $P(H_0)$ and $P(H_1)$ are prior probability. Because DFT coefficients are assumed to be independent, we have

$$\Lambda(m) = \left[\prod_{k=0}^{K-1} \frac{p(Y_k(m)|H_1)}{p(Y_k(m)|H_0)} \right]^{1/K} \tag{3}$$

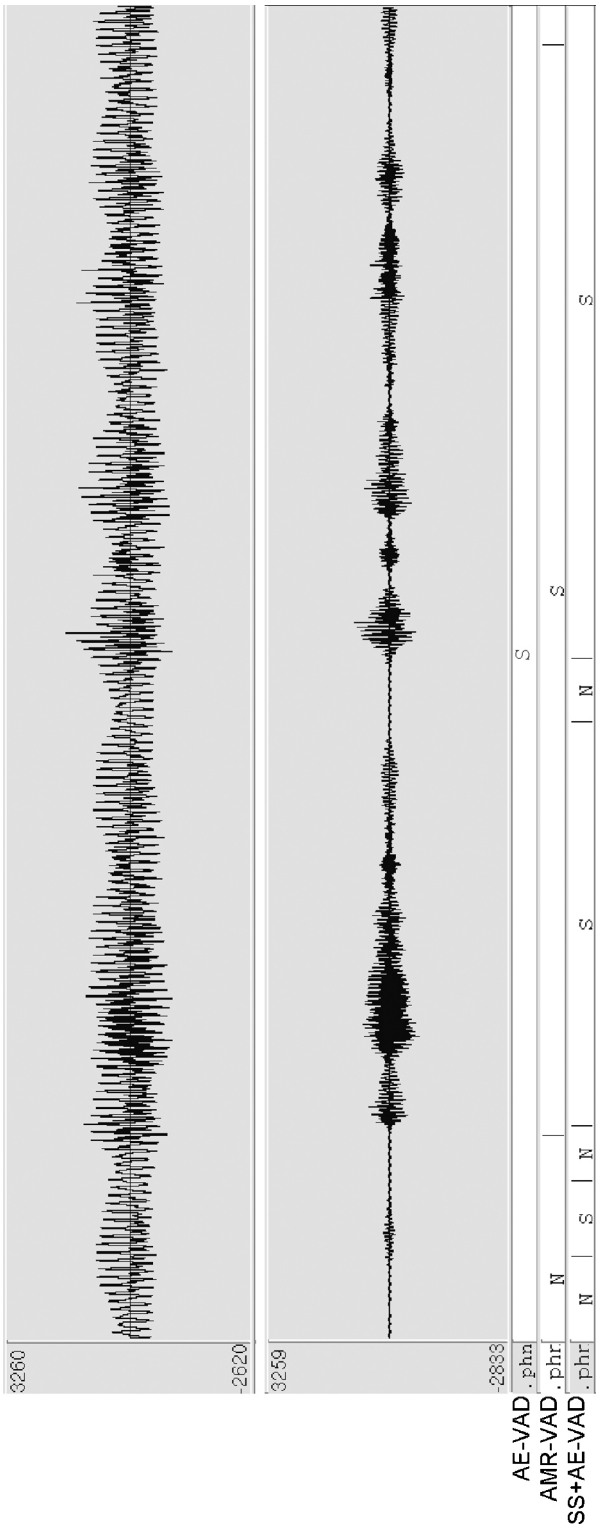


Fig. 4. *Top*: A short segment of low-energy interview speech in NIST 2008 SRE superimposed on periodic background noise. *Middle*: The same segment after spectral subtraction. *Bottom*: The VAD decisions (`S` for speech and `N` for non-speech). See Table 1 for the abbreviations of the VADs.

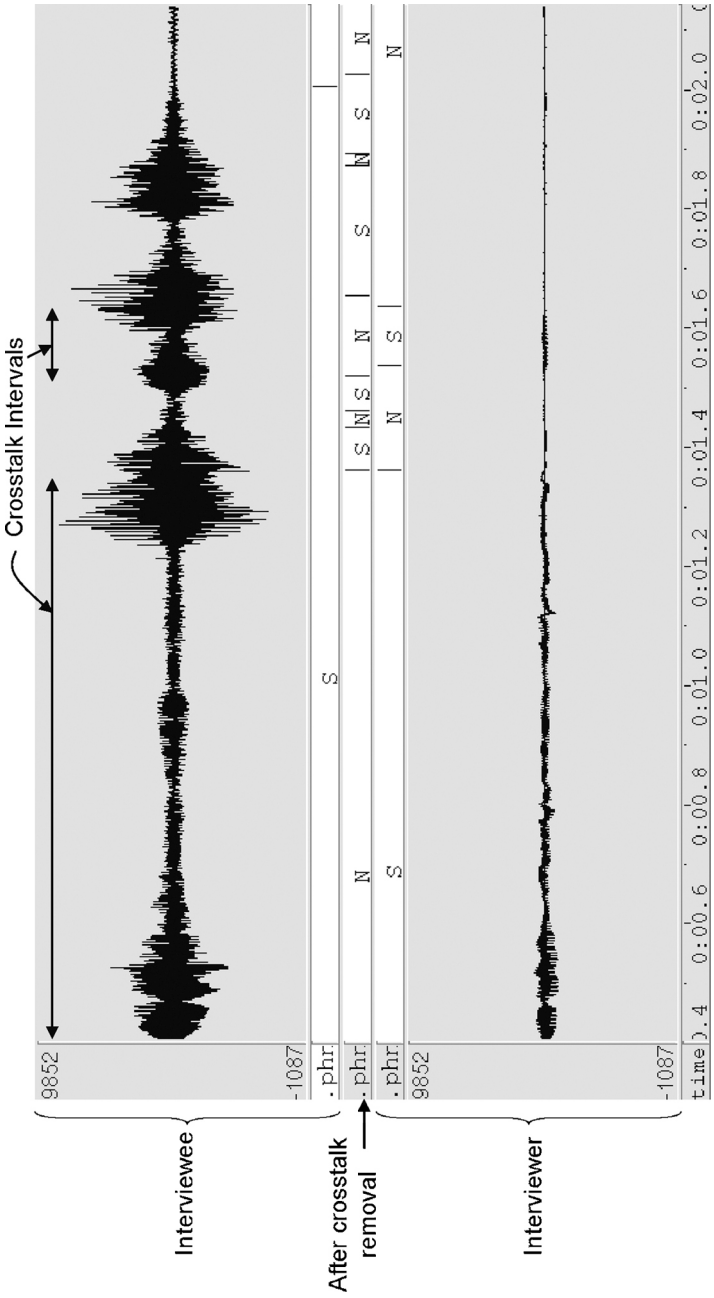


Fig. 5. The waveform of a short speech segment from an interviewee and interviewer respectively. The middle panel shows the corresponding VAD results (S for speech and N for non-speech) and the VAD decisions after considering the crosstalk.

where K is the number of frequency bins and $p()$'s are complex normal densities. The VAD score is then compared with a decision threshold η to make speech/non-speech decisions.

To apply SM-based VAD to detect speech segments in NIST SRE files, the SM scores $\Gamma(m)$ of the entire utterance are ranked in descending order as shown in Fig. 6. Then, a fixed percentage of scores in the lower and upper ends of the ranked list are selected and assumed to be the background frames and peak frames, respectively. The VAD's fixed decision threshold is a linear combination of the score mean of the lower end ($\overline{\Gamma}_b$) and the minimum score in the upper end:

$$\eta = \nu\overline{\Gamma}_b + (1-\nu)\min\{\Gamma(p_1), \ldots, \Gamma(p_L)\}, \tag{4}$$

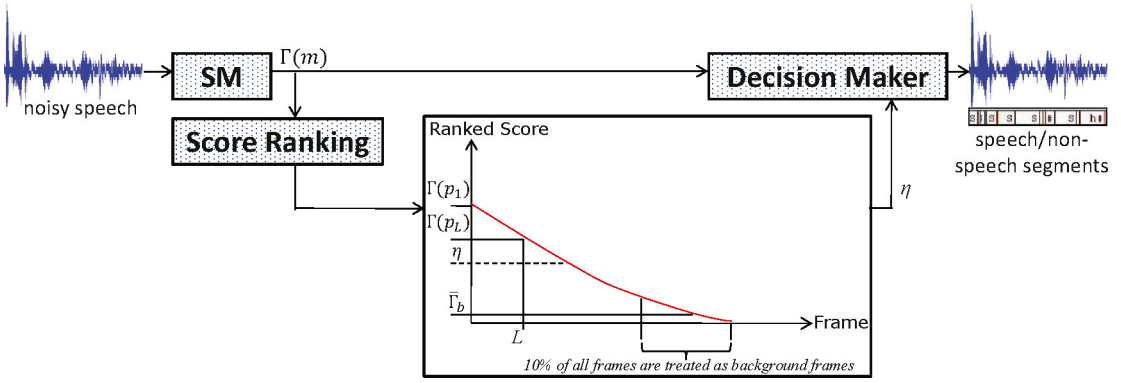


Fig. 6. The structure of SM-VAD incorporated with a fixed threshold.

where $0 \ll \nu < 1$ is a weighting factor and $\{\Gamma(p_1), \ldots, \Gamma(p_L)\}$ are top-L scores. Note that L cannot be too large; otherwise the rank list may include the peaks of some high-energy speech frames, which will lead to under-estimation of η. However, when L is too small, some medium-amplitude spikes will be missed. It was found that the influence of spikes can be largely eliminated by using the minimum amplitude in this ranked list, as evidenced by the VAD result in the Fig. 3.

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 assume that among all the frames in a speech file, 10% are background frames and 5% contain signal peaks.

3.2. Gaussian mixture model (GMM) based VAD

Mel-frequency cepstral coefficients (MFCCs) are known to be inadequate for discriminating speech and non-speech frames, primarily because of the similarity between the static MFCC vectors of speech and background noise. On the other hand, the harmonic structures of speech and background noise are more distinguishable and more noise robust (Gu and Rose, 2001). Based on this argument, Fukuda et al. (2010) extracted the harmonic-structure-based features from the middle range of the cepstral coefficients obtained from the discrete cosine transform (DCT) of the power spectral coefficients.

The cepstral coefficients $c_i(m)$ with small and large indexes i are liftered out because they include long and short oscillations. On the other hand, the coefficients in the middle part of the cepstrum capture the harmonic structure information in the human voice. The liftered cepstrum $\hat{c}_i(m)$ is converted back to the log power spectrum by inverse DCT, followed by the exponential transform to obtain the linear power spectrum. The power spectrum are finally converted to mel-cepstrum $\hat{q}_n(m)$ by applying a mel-scale filter bank and DCT, where n is the bin number of the harmonic structure-based mel cepstral coefficients. This feature captures the envelope information of the local peaks in the frequency spectrum corresponding to the harmonic information in the speech signals. Fig. 7 shows the procedure of extracting the harmonic-structure-based features.

Dynamic (spectro-temporal) features capture the variation of the spectral envelopes along the time axis. They are typically obtained by estimating the derivative of 5–9 consecutive acoustic vectors. The first-order derivative of a sequence of cepstral vectors is called delta cepstrum, and the second-order derivative is called delta-delta cepstrum.

In GMM-based VAD, the speech/non-speech decision at frame m is given by the log-likelihood ratio

$$\mathcal{L}(m) = \log p(\mathbf{y}(m)|H_1) - \log p(\mathbf{y}(m)|H_0) \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{5}$$

where the acoustic vectors $\mathbf{y}$'s are assumed to follow a mixture of Gaussian distribution:

$$p(\mathbf{y}|H_i) = \sum_{j=1}^{J} w_{ij} \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}) \tag{6}$$

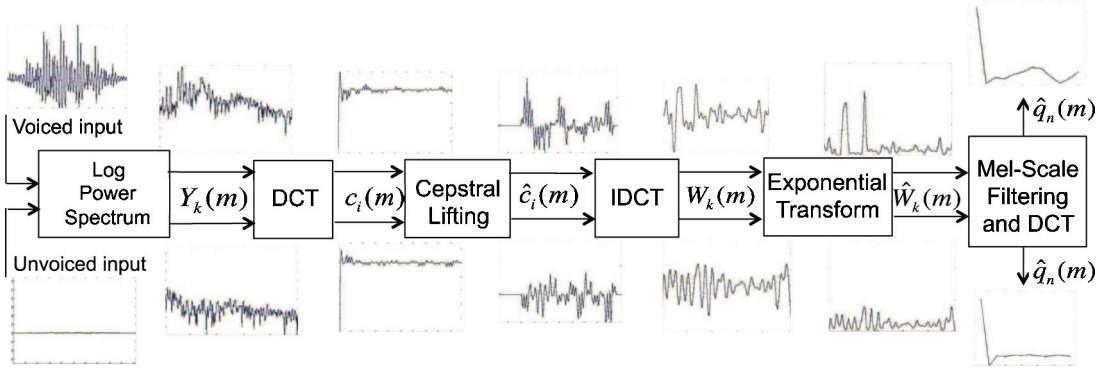


Fig. 7. The procedure of extracting the harmonic-structure-based features (after Fukuda et al., 2010).

where w_{ij}, $\boldsymbol{\mu}_{ij}$ and $\boldsymbol{\Sigma}_{ij}$ are the mixture weights, mean vectors and covariance matrices for either speech ($i=1$) or non-speech ($i=0$) model.

The decision threshold η is determined by a strategy similar to that of SM-VAD (Eq. (4)) described in Section 3.1. Specifically, 20% and 5% of a speech file are assumed to contain background frames and signal peaks, respectively.

Unlike the SM-based VAD, the GMM-based VAD requires the training of two GMMs – one representing speech and another one representing non-speech. This means that some speech files with speech and non-speech segmentations are required. In theory, the segmentations had better be the ground-truths, i.e., they need to be done by listening tests and human inspections of spectrograms. This is not a problem if clean speech files are available and the VAD is tested on the same files with noise added to them, e.g., the experiments in Fukuda et al. (2010). However, in NIST SREs, the requirement of ground-truth segmentations will cause difficulty because no clean speech files are available for the listening tests. Even if we can find some interview-style speech files with high enough SNR for the listening tests, they may be too clean and therefore cannot represent the realistic situations in other noisy speech files. Furthermore, hand labeling a large number of speech files is too laborious and time-consuming.

Here, we propose an automatic method that can find speech and non-speech segments that are close enough to the ground-truths for training the GMMs. Fig. 8 shows the procedure. Unlike the VAD in Fukuda et al. (2010), our GMM-based VAD contains an extra processing block – Frame Index Extraction – that finds the frame indexes of speech and non-speech segments with very high confidence of being correct. This seems to be a chicken-and-egg problem because if a reliable VAD exists, we do not need to build a new one in the first place. However, having some reliable speech and non-speech segments does not mean that we need a reliable VAD to detect both at the same time. The idea is that we can always make a simple energy-based VAD very reliable in detecting speech but extremely unreliable in detecting non-speech by adjusting the decision threshold such that it can achieve a very low false alarm (consider non-speech as speech) but having a very high missing rate (consider speech as non-speech). A similar argument applies to the reliable detection of non-speech. Because this simple VAD can only maintain either the false alarm or missing rate low but not both, it can only be used as a pre-processing step in more sophisticated VADs such as the one illustrated in Fig. 8.

The idea is to leverage the large number of speech files in NIST SREs. Specifically, for each interview-style speech files in the training set (e.g., past NIST SREs), a simple energy-based VAD is used to determine the energy of all

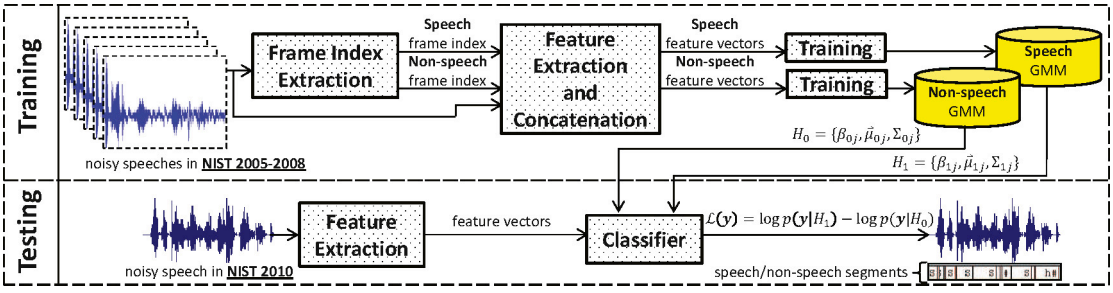


Fig. 8. Training and testing of the GMM-based VAD. See Fig. 9 for the algorithm of frame index extraction.

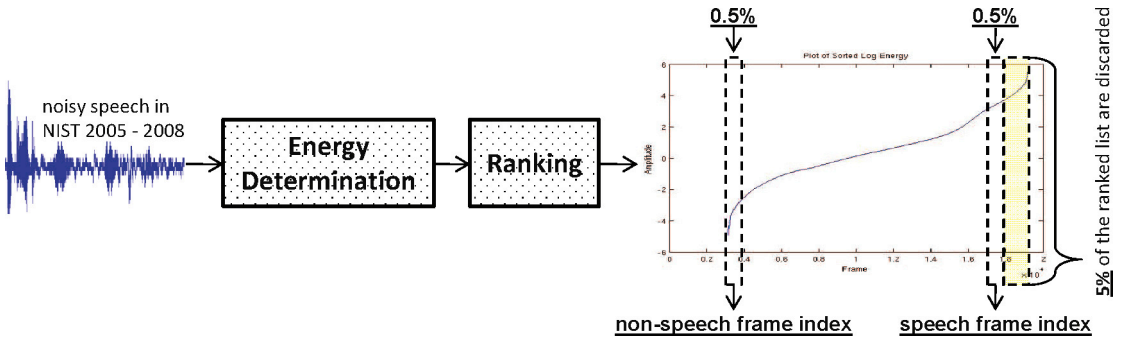


Fig. 9. The procedure of extracting the frame indexes representing the speech and non-speech segments in the processing block "Frame Index Extraction" in Fig. 8.

frames. Then, the frames are ranked in ascending order of energy as illustrated in Fig. 9. The top 5% of the ranked list are discarded because the high energy is most likely caused by spiky signals instead of speech. Because of the simplicity of the energy-based VAD, there will be many false alarms and misses in the detections. Therefore, only a small percentage in the upper- and lower-part of the ranked list are considered as speech and non-speech, respectively. More precisely, 99% of the frames in the middle of the ranked list will be discarded, and only the frames with a very high confidence of having a correct segmentation are retained for training the GMMs.

Given the frame indexes of speech and non-speech segments, static harmonic features and long-term dynamic features are extracted and concatenated, forming two streams of feature vectors as shown in Fig. 8. These concatenated features vectors are then used to train the GMMs. In this work, 3569 interview-style utterances from NIST SRE 2005–2008 were used for training the GMMs. This amount to 280,010 training vectors per GMM. The number of mixtures J was set to 32, and all Gaussians have a full covariance matrix.

One advantage of the GMM-based VAD is that it is less susceptible to spiky signals because these signals have low-level of harmonic contents and their temporal property is also different from that of speech signals. However, GMM-based VAD also has its own limitations. In particular, because the GMM-based VAD does not rely on SNR, it may falsely detect some weak cross-talks from other speakers as speech segments as long as the cross-talks contain speech-like characteristics. This drawback can be alleviated by using spectral subtraction as a pre-processor because the weak cross-talks will be considered as background signals so that they can be largely eliminated in the spectral subtraction process. Further discussions on the use of spectral subtraction as a pre-processor can be found in the next section.

4. Speech enhancement for VAD

Noise removal is a vital step for pre-processing the interview speech files in NIST SREs because many of them have low SNR. This paper proposes to apply spectral subtraction (SS) with a large over-subtraction factor to discard the background noise as much as possible before passing the enhanced speech to an energy-based VAD. Advanced speech enhancement techniques (e.g. MMSE (Ephraim and Malah, 1984) and LSA-MMSE (Ephraim and Malah, 1985)) have not been used because audio quality of reconstructed speech is not the main concern.[3] Instead, it is more important to increase the SNR in speech regions and to minimize the background noise in non-speech regions. Spectral subtraction meets this requirement without unnecessarily complicating the whole system.

4.1. Noise reduction via spectral subtraction

To obtain the enhanced speech $\hat{x}(t)$ from the noisy speech $y(t)$ at frame m, we implemented the spectral subtraction (Boll, 1979; Deller et al., 1993; Virag, 1999) of the form

$$\hat{X}_k(m) = \begin{cases} [|Y_k(m)| - \alpha(m)|\hat{B}_k|]e^{j\varphi_k(m)} & \text{if } |Y_k(m)| > (\alpha(m) + \beta(m))|\hat{B}_k| \\ \beta(m)|\hat{B}_k|e^{j\varphi_k(m)} & \text{otherwise,} \end{cases} \tag{7}$$

[3] Acoustic features (MFCCs) were extracted from the original signals instead of from the reconstructed signals.

Table 1
The voice activity detection (VAD) methods being applied in this paper and their acronym.

VAD		Description
1	AE-VAD	Energy-based VAD with the decision governed by the combination between average magnitude of background noise and signal peaks. The combination is controlled by a weighting factor (ν in Eq. (9))
2	ASR-VAD	Speech segments in the automatic speech recognition transcripts provided by NIST (Martin and Greenberg, 2010)
3	AMR-VAD	VAD in ETSI Adaptive Multi-Rate coder (Option2) (ETSI, 1999)
4	SM-VAD	Sohn's statistical-model-based VAD (Sohn et al., 1999) incorporated with a fixed threshold, determined by Eq. (4)
5	GMM-VAD	Gaussian-mixture-model-based VAD using long-term temporal information and harmonic structure-based features in noisy speech (Fukuda et al., 2010) incorporated with a fixed decision threshold
6	SS+SM-VAD	SM-VAD with spectral subtraction as a pre-processing step
7	SS+AE-VAD	AE-VAD with spectral subtraction as a pre-processing step

where k is the frequency bin index, $\varphi_k(m)$ is the phase of $Y_k(m)$, $\hat{B}_k$ is the average spectrum of some non-speech regions, $\alpha(m)$ 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 $\alpha(m)$ and $\beta(m)$ are computed as

$$\begin{aligned} \alpha(m) &= -\frac{1}{2}\gamma(m) + c \quad (\alpha_{\min} \le \alpha(m) \le \alpha_{\max}) \\ \beta(m) &= \begin{cases} \beta_{\min} & \text{if } \gamma(m) < 1 \\ \beta_{\max} & \text{otherwise} \end{cases} \end{aligned} \tag{8}$$

where $\gamma(m) = (\sum_k |Y_k(m)|)/(\sum_k |\hat{B}_k|)$ is the a posteriori SNR, c is a constant (=4.5 in this work), $\alpha_{\min}$, $\alpha_{\max}$, $\beta_{\min}$, and $\beta_{\max}$ constrain the allowable range of the over-subtraction factor and the noise floor. We set these values such that the speech spectra are over-subtracted when the SNR is low. In this work, we set $\alpha_{\max} = 4$, $\alpha_{\min} = 0.5$, $\beta_{\max} = 0.05$, and $\beta_{\min} = 0.01$. These values were determined empirically through experimentations on an i-vector systems (see Section 5.5) and by visual comparison between the original and reconstructed waveforms of several speech files. Because of the small $\beta_{\max}$ ($\ll 1$), musical noise occurs when some frequency components meet the condition in the upper branch of Eq. (7) while some others do not. While musical noise appears in both speech and non-speech regions, its energy in non-speech regions is not high enough to cause false detections, as evident in Fig. 1(d). Also, although this musical noise will degrade the perceptual quality of the denoised speech, it is not a concern here because the denoised speech is only used for VAD, not for speaker recognition, i.e., our goal is to detect voice activity rather than speech enhancement.

Note that if the background noise is high, consonants with weak energy will be masked by the noise. Therefore, it is more appropriate to exclude these weak consonants by means of over subtraction. On the other hand, if the background noise is low, $|\hat{B}_k|$ in Eq. (7) is almost zero, meaning that consonants will also be included.

Fig. 2 shows the histograms of the speech files in 2008 and 2010 SREs before and after spectral subtraction. Evidently, spectral subtraction can improve the SNR significantly.

4.2. Threshold determination and VAD decision logic

Fig. 10 shows the structure of the proposed energy-based VAD, which we refer to as SS+AE-VAD. For each utterance, after noise removal, the energy of each 10-ms frame is computed at every 1 ms. To avoid excessive fluctuation in the energy profile, a 40-tap moving average filter is applied to smooth the profile.

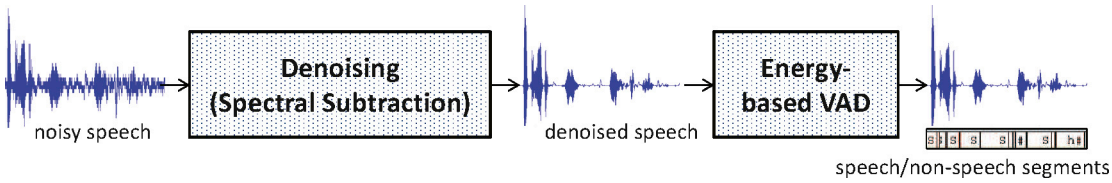


Fig. 10. The structure of the proposed VAD for NIST SREs.

The presence of spikes in some files affects the maximum SNR in these files, which 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. Consequently, low-energy speech segments could be mistakenly detected as non-speech. To address this problem, we have developed a similar strategy as the one in Section 3.1, but considering signal amplitude rather than statistical scores. The decision threshold is a linear combination of the mean of background amplitude ($\bar{a}_b$) and the minimum of the signal peaks:

$$\eta = \nu\bar{a}_b + (1-\nu)\min\{a(p_1), \ldots, a(p_L)\}, \tag{9}$$

where $\{a(p_1), \ldots, a(p_L)\}$ are the amplitudes (after the moving average filter) of L largest-amplitude frames. 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.

Fig. 1(b) and (d) shows the same speech file and segment as in Fig. 1(a) and (c) but after spectral subtraction. Evidently, with the background noise largely removed, speech and non-speech intervals can be correctly detected by an energy-based VAD. To highlight the advantage of spectral subtraction, Fig. 1(c) and (d) compares the segmentation results of SS+AE-VAD and that of the ETSI-AMR coder (Option 2). The figure suggests that this coder over-estimates the length of speech segments, whereas the SS+AE-VAD correctly detects the speech segments.

5. Experiments and results

VAD algorithms are typically evaluated by comparing the VAD decisions on clean speech against the VAD decisions on noise contaminated speech (Basbug et al., 1999), with performance shown on a receiver operating characteristic (ROC) curve. However, the noisy speech files in NIST SREs do not have their clean counterparts. Instead of hand labeling thousands of speech files, we used the performance indexes of speaker verification, i.e. equal error rate (EER), detection error tradeoff (DET) curves, and minimum normalized decision cost function (DCF) for quantifying VAD performance. Discussions and explanations of these performance indexes can be found in Martin et al. (1997) and the evaluation plans of NIST SRE.[4]

The experiments involve seven VADs, as shown in Table 1. Among the five conventional VADs (VADs 1–5), we applied spectral subtraction to the best performing (SM-VAD) and the worst performing (AE-VAD) ones, resulting in SS+SM-VAD and SS+AE-VAD in the last two rows of Table 1. By comparing the speaker verification performance obtained by these VADs against the ones without spectral subtraction, we can observe the contribution of spectral subtraction to the VAD performance.

5.1. Speech data

NIST 2005–2010 SREs were used in the experiments. NIST'05–08 SREs were used as development data, and NIST'10 was used for performance evaluations. Only male speakers in these corpora were used. The core task of NIST'10 is divided into nine common conditions. Conditions 1, 2, 4, 7 and 9 were considered because interview speech and telephone speech collected by different microphones are involved in these five conditions. Detail descriptions of these five conditions can be found in Section 4 of Martin and Greenberg (2010).

5.2. Speaker modeling and channel mismatch compensation

The target-speakers were modeled by GMM-SVM (Campbell et al., 2006) and i-vectors (Dehak et al., 2011).

For the GMM-SVM systems, we extracted 12 MFCCs (Davis and Mermelstein, 1980) and their first derivatives from the speech regions of the utterances to create 24-dim acoustic vectors. Cepstral mean normalization (Atal, 1974) was applied to the MFCCs, followed by feature warping (Pelecanos and Sridharan, 2001). A 512-center gender-dependent universal background model was created by using the interview utterances of NIST'05–06. MAP adaptation (Reynolds et al., 2000), 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 NIST'05–06) to create 300

[4] http://www.itl.nist.gov/iad/mig/tests/spk/2010/index.html.

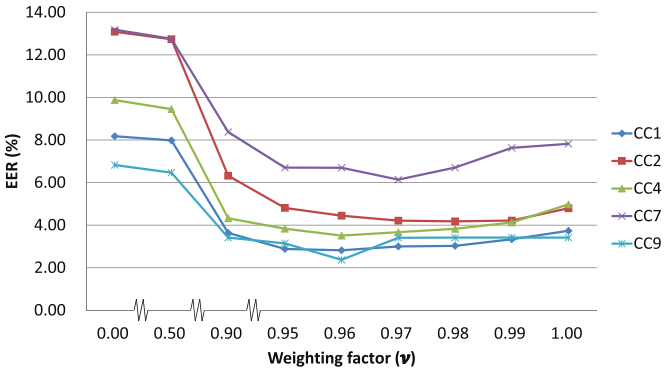


Fig. 11. Equal error rate of GMM-SVM against weighting factor ν in Eq. (9) under common conditions (CC) 1, 2, 4, 7 and 9 in NIST'10 (male). SS+AE-VAD (see Table 1) was used in all cases.

impostor GMMs. The mean vectors of these GMMs were stacked to produce 12288-dim GMM-supervectors (Campbell et al., 2006). Finally, a GMM-SVM speaker model for each target speaker is trained by using his target-dependent GMM-supervector and the background GMM-supervectors. The utterances of 144 male speakers from NIST'05–08 were used for estimating the gender-dependent nuisance attribute projection (NAP) matrices (Campbell et al., 2006) to reduce channel effects (NAP corank was set to 128). Each of these 144 speakers has at least 8 utterances. For the T-norm speaker models (Auckenthaler et al., 2000), 300 male utterances from NIST'05 were used.[5] The same set of background speakers used for creating the target-speaker SVMs were used for creating the T-norm SVMs.

For the i-vector systems, 19 MFCCs together with energy plus their first and second derivatives were extracted from the speech regions as detected by the VADs, followed by cepstral mean normalization and feature warping with a window size of 3 s. A 60-dim acoustic vector was extracted every 10 ms, using a Hamming window of 25 ms. NIST 2006–2008 microphone data were used to train a 1024-center UBM. We selected 6102 utterances from 191 speakers in NIST 2005–2008 SREs to estimate a total variability matrix with 400 total factors. Then, we used 9511 utterances of these 191 speakers to estimate the Gaussian PLDA (Garcia-Romero and Espy-Wilson, 2011) loading matrix with 150 latent variables. We applied length normalization (Garcia-Romero and Espy-Wilson, 2011) to all i-vectors before computing the loading matrix.

5.3. Selection of threshold parameters for SS+AE-VAD

As mentioned in Section 4, energy-based VAD requires a decision threshold for making speech/non-speech decisions. An experiment was conducted to investigate the effect of varying the weighting factor ν (Eq. (9)) on the energy-based VAD.

Fig. 11 suggests that the best range of ν in Eq. (9) is between 0.95 and 0.99. Once this value drops below 0.95, the performance degrades rapidly. This implies that the peak amplitudes can only be used as a reference for setting the VAD decision threshold, whereas the background amplitudes are more trustworthy. However, the threshold cannot totally relies on the background amplitude, because the EER and minDCF increase when ν increases from 0.99 to 1.0.

5.4. Comparing different VADs

based on the results in Section 5.3, the weighting factor ν in Eq. (9) was set to 0.95 and 0.96 for AE-VAD and SS+AE-VAD, respectively. For SM-VAD, SS+SM-VAD, and GMM-VAD, ν in Eq. (4) was set to 0.993.

Table 2 shows the equal error rate (EER) and minimum normalized decision cost function (minNDCF) achieved by the GMM-SVM systems. The results suggest that preprocessing the noisy sound files by spectral subtraction is a promising idea. After applying SS, the AE-VAD and SM-VAD reduce the overall EER by 56% and 5% respectively.

[5] It was drawn to the authors' attention that performance could be improved if impostors and T-norm speakers from NIST'08 were used because they match the test speakers in NIST'10 better.

Table 2
Performance of GMM-SVM systems based on 7 VADs under common conditions (CC) 1, 2, 4, 7 and 9 of NIST 2010 SRE (male speakers). Refer to Table 1 for the definition of the VADs.

VAD method	Equal error rate (%)						Minimum normalized DCF					
	CC1	CC2	CC4	CC7	CC9	Overall	CC1	CC2	CC4	CC7	CC9	Overall
AE-VAD	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-VAD	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-VAD	4.44	8.05	9.44	12.85	5.98	9.61	0.81	0.85	0.80	0.77	0.55	0.90
GMM-VAD	3.64	5.68	5.71	8.93	4.27	6.28	0.71	0.72	0.72	0.63	0.45	0.82
SM-VAD	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-VAD	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-VAD	**2.82**	**4.44**	**3.51**	**6.70**	**2.37**	**4.55**	0.70	**0.58**	**0.62**	0.64	**0.17**	**0.72**

Each boldface number represents the best performance under the respective common condition.

Fig. 12 shows the DET performance achieved by the seven VADs. The results show that SS+AE-VAD achieves a significant lower error rates than the ETSI-AMR coder, ASR transcripts and the simple energy-based VAD for a wide range of operating points.

The results shows that SM-VAD performs better than GMM-VAD for detecting interview speech in NIST SREs. Note that this result does not mean that SM-VAD is better than GMM-VAD for all tasks. In fact, the conditions in NIST SREs are disadvantageous to GMM-VAD because the GMMs of this VAD requires a large number of ground-truth speech and non-speech segments for training. Unfortunately, in NIST SREs, such segments are not available. Therefore, we developed an automatic approach (see Section 3.2) to finding a large number of speech and non-speech segments that are close enough to the ground-truth as substitutions. The SM-VAD, on the other hand, does not require any ground-truth segments. Therefore, it is more appropriate for NIST SREs.

We notice that both SS and SM work well for the interview speech in NIST 2010 SRE. The error rates achieved by SS+AE-VAD, however, are slightly lower than that achieved by SM-VAD.

Comparing the results of AE-VAD and SS+AE-VAD reveals that SS has significant contribution to the conventional energy-based VAD. However, the performance of SS+SM-VAD is only slightly better than that of SM-VAD. This suggests that SS is not vital 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. As pre-processing the noisy speech by spectral

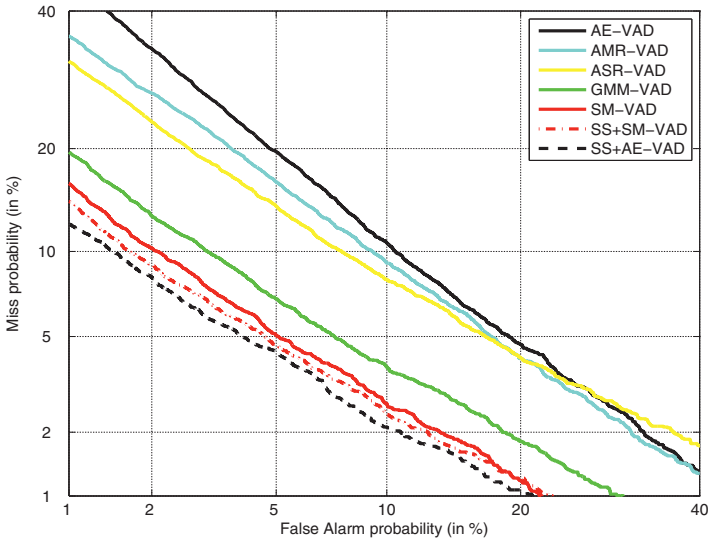


Fig. 12. DET performance of GMM-SVM systems for all trials involving interview-style speech in NIST'10 (male). Labels in the legend are arranged in descending EER. Refer to Table 1 for the definitions of different VADs in the legend.

subtraction is another approach to using the background spectrum, therefore in SS+SM-VAD, the background spectrum has been used twice. As a result, the gain of applying SS to SM-VAD is not as significant as applying SS to AE-VAD.

Note that SS+AE-VAD and SM-VAD use the background spectrum in a different manner. 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-VAD. Specifically, to remove as much background noise as possible, we may apply a large upper-limit for the over-subtraction factor ($\alpha_{\max}$) and a small lower-limit for the noise floor ($\beta_{\min}$). The over-subtraction factor $\alpha(m)$ 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 in low SNR region whereas more speech content will be retained in high SNR region. The SM-VAD, 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 transcripts do 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. To ensure the intelligibility of the encoded speech, it is important for the VAD in a speech coder to include speech onsets and offsets. However, this overestimation is not appropriate for speaker verification, as excessive amount of non-speech will be used for verification.

5.5. Performance of I-vector systems

While Table 2 suggests that spectral subtraction is an appropriate pre-processing step for both energy-based VAD and statistical-model based VAD, it is of interest to investigate if spectral subtraction is also suitable for state-of-the-art i-vector systems. To further compare the performance of the proposed VAD with more advanced statistical VADs, the distribution of DFT coefficients in SM-VAD were assumed to follow not only Gaussian distributions but also Laplacian and Gamma distributions, similar to that of Chang et al. (2006).

The performance of SM-VAD (Gaussian) with and without spectral subtraction is shown in Table 3. Evidently, spectral subtraction can help the SM-VAD. Results also show that SM-VAD based on Gamma distributions performs slightly better than Sohn's classical SM-VAD in terms of EER, but in terms of minimum DCF, Sohn's SM-VAD peforms better.

Recall from Eq. (8) that there are several parameters in spectral subtraction. It is of interest to see if the performance of SS+AE-VAD is sensitive to these parameters. Among the parameters in this VAD, $\alpha_{\max}$ and $\beta_{\max}$ have the greatest impact on the denoised waveform. Therefore, we varied these two parameters and investigated how they affect the performance of the i-vector systems. As shown in Table 3, the performance of SS+AE-VAD is the best when $\alpha_{\max} = 4$ or 8 and $\beta_{\max} = 0.05$, which also agree with the configuration we used for the GMM-SVM system in Table 2.

Table 3
Performance of an i-vector system based on statistical-model based VAD (SM-VAD) with three different distributions of DFT coefficients and spectral-subtraction VAD (SS+AE-VAD and SS+SM-VAD) under the interview-interview conditions of NIST 2010 SRE.

Method	Equal error rate (%)			Minimum normalized DCF		
	CC1	CC2	CC1 &2	CC1	CC2	CC1 &2
SM-VAD (Gaussian)	1.82	3.14	3.02	**0.295**	0.468	0.487
SM-VAD (Laplacian)	2.02	2.90	2.92	0.299	0.485	0.536
SM-VAD (Gamma)	1.92	2.91	2.87	0.314	0.487	0.540
SS+SM-VAD (Gaussian)	1.48	**2.31**	**2.30**	0.315	0.483	0.511
SS+AE-VAD ($\beta_{\max} = 0.05$)						
$\alpha_{\max} = 1$	2.93	4.76	4.43	0.395	0.665	0.617
$\alpha_{\max} = 2$	**1.40**	2.59	2.54	0.353	0.466	0.512
$\alpha_{\max} = 4$	1.60	2.68	2.55	0.351	**0.463**	**0.470**
$\alpha_{\max} = 8$	1.60	2.68	2.55	0.351	**0.463**	**0.470**
SS+AE-VAD ($\alpha_{\max} = 4$)						
$\beta_{\max} = 0.1$	1.69	2.68	2.55	0.356	0.458	0.481
$\beta_{\max} = 0.5$	4.35	4.93	4.92	0.354	0.493	0.518

As $\alpha_{\max}$ determines the maximum amount of noise to be subtracted from the noisy signal, the good performance at large value of $\alpha_{\max}$ in Table 3 confirms our earlier argument in Section 4.1 that it is desirable to use over-subtraction to remove as much noise as possible when the SNR is low.

6. Conclusions and future work

A voice activity detector specially designed for extracting speech segments from the interview-speech files in NIST SREs has been proposed and evaluated under the NIST 2010 SREs protocols. Several conclusions can be drawn from this work: (1) noise reduction is of primary importance for energy-based VAD under low SNR; (2) it is important to remove the sinusoidal background noise as this kind of background signal could lead to many false detection in energy-based VAD; (3) a reliable threshold strategy is required to address the spiky (impulsive) speech signals; and (4) our proposed spectral subtraction VAD outperforms the segmentations derived from the ASR transcripts provided by NIST, the VAD in the advanced speech coder (ETSI-AMR, Option2), the state-of-the-art statistical-model-based VAD, and Gaussian-mixture-model-based VAD in speaker verification.

The proposed VAD is optimized for interview speech in NIST SREs. It is of interest to investigate its performance on other databases, including those in speech recognition such as CENSREC-1-C speech database (Kitaoka et al., 2007) and Aurora 2 database (Hirsch and Pearce, 2000). The present study assumes that background noise is stationary. It is of interest to apply methods – such as Ghosh et al. (2011) – that can deal with non-stationary noises in NIST SREs.

Acknowledgement

This work was in part supported by The Hong Kong Research Grant Council, Grant No. PolyU5264/09E and PolyU Grant No. G-YL78.

References

Atal, B.S., 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. Journal of the Acoustical Society of America 55 (6), 1304–1312.

Auckenthaler, R., Carey, M., Lloyd-Thomas, H., 2000. Score normalization for text-independent speaker verification systems. Digital Signal Processing 10 (1–3), 42–54.

Basbug, F., Nandkumar, S., Swaminathan, K., 1999. Robust voice activity detection for DTX operation of speech coders. In: IEEE Workshop on Speech Coding, pp. 58–60.

Benyassine, A., Shlomot, E., Su, H.Y., 1997 September. ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications. International Telecommunication Union.

Beritelli, F., Casale, S., Ruggeri, G., Serrano, S., 2002. Performance evaluation and comparison of G.729/AMR/Fuzzy voice activity detectors. IEEE Signal Processing Letters 9 (3), 85–88.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP 27 (2), 113–120.

Campbell, W.M., Sturim, D.E., Reynolds, D.A., 2006. Support vector machines using GMM supervectors for speaker verification. IEEE Signal Processing Letters 13 (5), 308–311.

Campbell, W.M., Sturim, D.E., Reynolds, D.A., Solomonoff, A., 2006. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In: Proceeding of ICASSP, Vol. 1, Toulouse, France, pp. 97–100.

Chang, J.H., Kim, N.S., Mitra, S.K., 2006. Voice activity detection based on multiple statistical models. IEEE Transactions on Signal Processing 54 (6), 1965–1976.

Chengalvarayan, R., 1999. Robust energy normalization using speech/non-speech discriminator for German connected digit recognition. In: Proc. EUROSPEECH 1999, Budapest, Hungary, pp. 61–64.

Cornu, E., Sheikhzadeh, H., Brennan, R.L., Abutalebi, H.R., Tam, E.C.Y., Iles, P., Wong, K.W., 2003. ETSI AMR-2 VAD: evaluation and ultra low-resource implementation. In: Acoustics, Speech, and Signal Processing (ICASSP'03), Hong Kong, pp. 585–588.

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

Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on ASSP 28 (4), 357–366.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. IEEE Trans. on Audio, Speech, and Language Processing 19 (4), 788–798.

Deller Jr, J.R., Proakis, J.G., Hansen, J.H.L., 1993. Discrete-time Processing of Speech Signals. Macmillan Publishing Company, Upper Saddle River, New Jersey.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech and Signal Processing 32 (6), 1109–1121.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Transactions on Acoustics, Speech and Signal Processing 33, 443–445.

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

ETSI, 2001. Universal Mobile Telecommunication Systems (UMTS); Mandatory Speech Codec speech processing functions, AMR speech codec; Voice Activity Detector VAD, etsi ts 126 094 v4.00 Edition (2001-03).

Freeman, D.K., Cosier, G., Southcott, C.B., Boyd, I., 1989. The voice activity detector for the Pan-European digital cellular mobile telephone service. In: Acoustics, Speech, and Signal Processing, 1989. ICASSP-89, Vol. 1, Glasgow, UK, pp. 369–372.

Fukuda, T., Ichikawa, O., Nishimura, M., 2010. Long-term spectro-temporal and static harmonic features for voice activity detection. IEEE Journal of Selected Topics in Signal Processing 4 (5), 834–844.

Góriz, J.M., Ramíez, J., Lang, E.W., Puntonet, C.G., Turias, I., 2010. Improved likelihood ratio test based voice activity detector applied to speech recognition. Speech Communication 52 (7–8), 664–677.

Garcia-Romero, D., Espy-Wilson, C., 2011. Analysis of I-vector length normalization in speaker recognition systems. In: Interspeech'2011, pp. 249–252.

Ghosh, P., Tsiartas, A., Narayanan, S., 2011. Robust voice activity detection using long-term signal variability. IEEE Transactions on Audio, Speech, and Language Processing 19 (3), 600–613.

Gu, L., Rose, K., 2001. Perceptual harmonic cepstral coefficients for speech recognition in noisy environment. In: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings (ICASSP '01), Vol. 1, Salt Lake City, UT, USA, pp. 125–128.

Hautamaki, V., Tuononen, M., Niemi-Laitinen, T., Franti, P., 2007. Improving speaker verification by periodicity based voice activity detection. In: Proceedings of the 12th International Conference on Speech and Computer (SPECOM'2007), Vol. 2, Moscow, pp. 645–650.

Hirsch, H., Pearce, D., 2000. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW), Paris, pp. 181–188.

Kinnunen, T., Saastamoinen, J., Hautamaki, V., Vinni, M., Franti, P., 2009. Comparing maximum *a posteriori* vector quantization and Gaussian mixture models in speaker verification. In: Acoustics, Speech and Signal Processing, 2009, ICASSP 2009, Taipei, pp. 4229–4232.

Kitaoka, N., Yamamoto, K., Kusamizu, T., Nakagawa, S., Yamada, T., Tsuge, S., Miyajima, C., Nishiura, T., Nakayama, M., Denda, Y., et al., 2007. Development of VAD evaluation framework CENSREC-1-C and investigation of relationship between VAD and speech recognition performance. In: Automatic Speech Recognition & Understanding, 2007, ASRU IEEE Workshop on, pp. 607–612.

Li, Q., Zheng, J.S., Tsai, A., Zhou, Q., 2002. Robust endpoint detection and energy normalization for real-time speech and speaker recognition. Speech and Audio Processing, IEEE Transactions on 10 (3), 146–157.

Mak, M.W., Yu, H.B., 2010. Robust voice activity detection for interview speech in nist speaker recognition evaluation. In: Proceedings of the APSIPA ASC 2010, Singapore.

Marciniak, T., Rochowniak, R., Dabrowski, A., 2008. Subband wavelet signal denoising for voice activity detection. In: Signal Processing Algorithms, Architectures, Arrangements, and Applications (SPA), 2008, pp. 93–96.

Martin, A.F., Greenberg, C.S., 2010. The NIST 2010 speaker recognition evaluation. In: Interspeech, Japan, pp. 2726–2729.

Martin, A., Greenberg, C. (Eds.), 2010. NIST SRE10 workshop, NIST Multimodal Information Group. Brno, Czech Republic.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., 1997. The DET curve in assessment of detection task performance. In: Proceedings of Eurospeech'97, pp. 1895–1898.

Marzinzik, M., Kollmeier, B., 2002. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. IEEE Transactions on Speech and Audio Processing 10 (2), 109–118.

Nemer, E., Goubran, R., Mahmoud, S., 2001. Robust voice activity detection using higher-order statistics in the LPC residual domain. IEEE Transcations on Speech and Audio Processing 9 (3), 217–231.

Pelecanos, J., Sridharan, S., 2001. Feature warping for robust speaker verification. In: Proceedings of Odyssey: The Speaker and Language Recognition Workshop, Crete, Greece, pp. 213–218.

Ramirez, J., Segura, J., Benitez, C., de La Torre, A., Rubio, A., 2004. Voice activity detection with noise reduction and long-term spectral divergence estimation. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings (ICASSP'04), Vol. 2, pp. 1094–1097.

Ramirez, J., Segura, J.C., Gorriz, J.M., Garcia, L., 2007. Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition. IEEE Transactions on Audio Speech Language Processing 15 (8), 2177–2189.

Ramirez, J., Gorriz, J.M., Segura, J.C., 2007. Robust speech recognition and understanding, I-Tech, Vienna, Austria, Ch. Voice activity detection. In: Fundamentals and Speech Recognition System Robustness, pp. 1–22.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10 (1–3), 19–41.

Sangwan, A., Chiranth, M.C., Jamadagni, H.S., Sah, R., Prasad, R.V., Gaurav, V., 2002. VAD techniques for real-time speech transmission on the internet. In: IEEE International Conference on High-Speed Networks and Multimedia Communications, pp. 46–50.

Sohn, J., Kim, N.S., Sung, W., 1999. A statistical model-based voice activity detection. IEEE Signal Processing Letters 6 (1), 1–3.

Sun, H., Ma, B., Li, H., 2008. An efficient feature selection method for speaker recognition. In: ISCSLP'08, pp. 1–4.

Sun, H., Nwe, T., Ma, B., Li, H., 2009. Speaker diarization for meeting room audio. In: Proceedings of Interspeech, pp. 900–903.

Tanyer, S.G., Ozer, H., 2000. Voice activity detection in nonstationary noise. IEEE Transcations on Speech and Audio Processing 8 (4), 478–482.

Torre, A.D.L., Ramirez, J., Benitez, C., Segura, J.C., Garcia, L., Rubio, A.J., 2006. Noise robust model-based voice activity detection. In: Interspeech, Pittsburgh, Pennsylvania, pp. 1954–1957.

Tucker, R., 1992. Voice activity detection using a periodicity measure. Communications, Speech and Vision, IEE Proceedings I 139 (4), 377–380.

Varela, Ó., San-Segundo, R., Hernández, L., 2011. Combining pulse-based features for rejecting far-field speech in a HMM-based voice activity detector. Computers & Electrical Engineering 37 (4), 589–600.

Virag, N., 1999. Single channel speech enhancement based on masking properties of the human auditory system. IEEE Transactions on Speech and Audio Processing 7 (2), 126–137.

Vlaj, D., Kacic, Z., Kos, M., 2012. Voice activity detection algorithm using nonlinear spectral weights, hangover and hangbefore criteria. Computers & Electrical Engineering 38 (6), 1820–1836.

Woo, K.H., Yang, T.Y., Park, K.J., Lee, C.Y., 2000. Robust voice activity detection algorithm for estimating noise spectrum. Electronics Letters 36 (2), 180–181.

Yu, H., Mak, M., 2011. Comparison of voice activity detectors for interview speech in NIST speaker recognition evaluation. In: Interspeech, pp. 2353–2356.