Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

Supervised/Unsupervised Voice Activity Detectors for Text-dependent Speaker Recognition on the RSR2015 Corpus

Md Jahangir Alam, Patrick Kenny, Pierre Ouellet, Themos Stafylakis, Pierre Dumouchel

[1]Centre de recherche informatique de Montréal, Montréal, Canada
[2]École de technologie supérieure, Montréal, Canada
{jahangir.alam, patrick.kenny, pierre.ouellet, themos.stafylakis}@crim.ca

Abstract

Voice activity detection, i.e., discrimination of the speech/non-speech segments in a speech signal, is an important enabling technology for a variety of speech-based applications including the speaker recognition. In this work we provide a performance evaluation of the following supervised and unsupervised VAD algorithms in the context of text-dependent speaker recognition on the RSR2015 (Robust Speaker Recognition 2015) task : Energy-based VAD with and without hangover scheme and endpoint detection, vector quantization-based VAD, Gaussian mixtures model (GMM)-based VAD (both supervised and unsupervised way), and sequential GMM-based VAD. Experimental results show that both the supervised and unsupervised GMM-based VADs perform better than the other VAD algorithms. Considering all three evaluation metrics (equal error rate, old (SRE 2008) and new (SRE 2010) normalized detection cost functions) unsupervised GMM-based VAD performed the best.

10.21437/Odyssey.2014-14

1. Introduction

Voice activity detection (VAD) is a fundamental task in various speech-related applications, such as speech coding, speech enhancement, speaker diarization, speaker and speech recognition. It is often defined as the problem of distinguishing active speech from nonspeech (silence and/or noise) in a utterance. One major step which affects directly the performance of speaker/speech recognition systems is the detection of speech from audio stream. For example, too many false alarms, or too many nonspeech segments wrongly detected as speech and used in the training can corrupt the acoustic models, and hence reduces recognition accuracy. On the other hand, during testing, if not enough speech segments are detected then the speaker/speech recognition algorithms will not be able to detect the speaker/full spoken sentence. Therefore, accurate determination of active speech from nonspeech in a recording both in clean and noisy environments is an important task. Depending on the surrounding environment of the recording, nonspeech can be silence, noise, music, or a variety of other acoustical signals such as door knocking, coughing, paper shuffling, heating ventilation and air conditioning, passing of a vehicle, train, or even background speech [1]. One of the main components in any VAD algorithm is the extraction of relevant features such as energy, signal-to-noise ratio (SNR), periodicity, dynamics of speech, zero crossing rate (ZCR) from the given recording that can represent discriminative characteristics of speech comparing to nonspeech. More recent VAD algorithms, while utilizing the many of the same features, use statistical models to distinguish speech/nonspeech based on the average of the log-likelihood ratios between the observed signal and background noise in individual frequency bins [2]. In [3] contextual information derived from multiple observations has been incorporated into the likelihood ratio tests (LRT) and a novel way to improve the robustness of existing LRT-based VADs has been proposed in [5] by selecting the harmonic frequency components for computing the likelihood ratio (LR) scores of the voiced frames.

In NIST Speaker Recognition (text-independent) Evaluations (SREs) participating sites typically used Energy-based VAD with/without spectral subtraction technique as a pre-processor [4, 8], phoneme recognizer-based VAD with a post-processing using short-term energy [6], ASR transcripts provided by NIST in the VAD [21], supervised Gaussian mixture models (GMM)-based VAD [7, 14].

For supervised VAD algorithm it can be difficult and laborious to obtain suitable training data. Therefore, it is desirable to design a VAD that is both robust and unsupervised, i.e., does not require a specialized training data. Recently there has been interest in developing unsupervised VAD algorithms that have the performance advantages of supervised techniques [11]. Some recently proposed unsupervised VADs are: vector quantization-based self-adaptive VAD [9], and sequential GMM-based VAD [10].

In this work we use unsupervised and supervised VAD algorithms for text-dependent speaker recognition task on the RSR2015 corpus [18]. Contrary to the text-independent speaker verification, a process of verifying the identity without constraint on the speech content, text-dependent speaker verification requires the speaker uttering the enrolled pass-phrase. The pass-phrase may be unique or user dependent or prompted by the system. The following VADs are considered: energy-based VAD [13], energy-based VAD with a hangover scheme, vector quantization-based VAD [9], sequential GMM-based VAD [10], supervised GMM-based VAD [6, 14]. We also use an unsupervised GMM-based VAD by combining the energy-based and log likelihood ratio (LLR)-based voice activity detection criteria, where the LLR is calculated using 16-component speech and non-speech GMMs [9, 12]. In order to train GMMs speech and nonspeech feature frames are separated from the observed signal based on a fraction of the lowest and highest energy frames [9].

2. Voice Activity Detectors

The voice activity detection (VAD) problem considers detecting the presence of speech in an utterance. A VAD usually has the following three modules [1]:

1. Feature extraction: The objective of this module is to extract discriminative features from the observed signal for detection.

2. Decision making: This module defines the rule or method for assigning a class (speech or nonspeech) based on the extracted feature.
3. Hangover scheme: This module, which is often implemented as a finite state machine, is employed to increase detection hits and reduce false alarms. The motivation for this module is found in the speech production process and the reduced signal energy of word beginnings and endings.
VAD algorithms considered in this work for performance evaluation in the context of text-dependent speaker verification task are described in this section.

2.1. Energy-based VAD

Energy is a simple measure of loudness of a signal. In the VAD literature energy is one of the most widely used features due to its simplicity and adequate performance in clean environment. The energy of the *m*th frame of a signal $s(m,n)$ is given by

$$E^{\log}(m) = 10\log_{10}\left(\frac{1}{N}\sum_{n=1}^{N} s(m,n)^2\right), \quad (1)$$

where n is the sample index, N is the frame length.
In this work we use two energy-based VAD:
***Energy-based VAD*:** CRIM's VAD software a slight modification from the software available from ISIP at Mississippi State University [13].
***Energy-based VAD I*:** This VAD is shown in fig. 1.

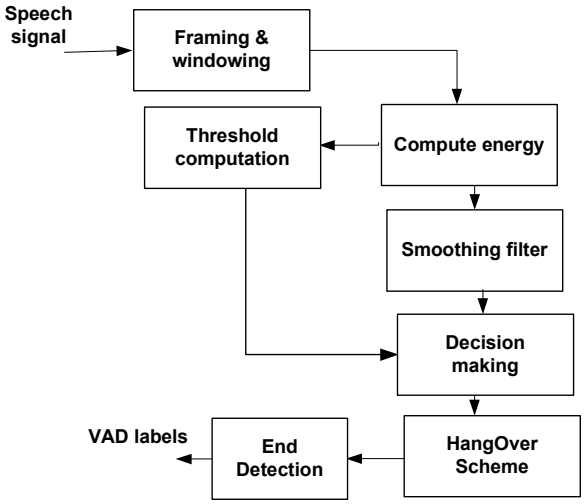

Figure 1:Energy-based voice activity detector with hangover scheme and end detection.

After computing the frame-wise energy of the speech signal a moving averaging filter is used with a 9-frames sliding window to smooth the decision boundaries. The decision threshold θ is then computed from the sorted $E^{\log}$ ($sE^{\log}$) using following formula:

$$\theta = \frac{\theta_1 + \theta_2}{2} \quad (2)$$

where θ_1 and θ_2 represent those values of sorted energy that correspond to the 20% and 80% length (or indices) of the sorted energy vector. VAD decision is made by comparing the energy of each frame to the decision threshold θ. Speech is active if the $E^{\log} > \theta$ otherwise non-speech is decided.
The VAD decision is then smoothed using a hang-over scheme. Most of the VAD algorithms that formulate the decision rule on a frame by frame basis normally use decision smoothing algorithms in order to improve the robustness against the noise. The motivations for these approaches are found in the speech production process and the reduced signal energy of word beginnings and endings. The so called hang-over scheme extends and smoothes the VAD decision in order to recover speech periods that are masked by acoustic noise. The hang over scheme influences the behavior of the VAD in a two distinct ways. Firstly the scheme delays the transition from the noise state to the speech state. This is done in such a way that if the VAD decision making process indicates speech then the final VAD decision is always speech. The delay is introduced to ensure the hangover scheme does not move into the speech state as a result of a false-alarm. The scheme secondly delays the transition from the speech state to the noise state, i.e., even if the VAD results indicates noise, the VAD will not necessarily decide noise, but will begin to progress through the transition states to the noise state. This effectively delays the transition from the speech state to the noise state and results in a reduction in miss detections. The VAD is thus quick to react to a change from noise to speech, but is slow to react to a change from speech to noise.
To get the final VAD labels we use an end-point detection algorithm that looks for the beginning and end of speech. It usually checks for a specified duration of silences in the VAD decision and if the silence duration is longer than that specific duration then it is considered to be out of the sentence. The outputs of end detector are frame indices that contain speech.

2.2. Vector Quantization (VQ)-based VAD [9]

Speech and non-speech segmentations were performed using an unsupervised voice activity detector proposed in [1]. Various steps of this VAD is shown in figure 2. At first log energy $\boldsymbol{E}^{\log}$ is computed for each frame after enhancing the speech signal using spectral subtraction technique. The goal of speech enhancement is to increase energy contrast between the speech and non-speech. The log energy values are sorted in ascending order. The lowest and highest energy frames (e.g., 10% of all frames in each case) are considered as non-speech and speech frames, respectively. 12-dimensional Mel-frequency cepstral coefficients (MFCCs) features are computed from the original signal (without speech enhancement). Using *k-means* (k = 16) clustering speech and non-speech models are then trained taking the MFCCs corresponding to the lowest and highest energy frame indices. If $\boldsymbol{x}_k^s$ and $\boldsymbol{x}_k^{ns}$ represent codevectors for the speech and non-speech, respectively, obtained using *k*-means and $\boldsymbol{c}_t$ represents the cepstral feature vector of *t*th frame then the Euclidian distance measures between $\boldsymbol{c}_t$ & $\boldsymbol{x}_k^s$ and $\boldsymbol{c}_t$ & $\boldsymbol{x}_k^{ns}$ are given by:

$$\boldsymbol{D}^s = \left\|\boldsymbol{c}_t - \boldsymbol{x}_k^s\right\|^2 \quad (3)$$

$$\boldsymbol{D}^{ns} = \left\|\boldsymbol{c}_t - \boldsymbol{x}_k^{ns}\right\|^2 \quad (4)$$

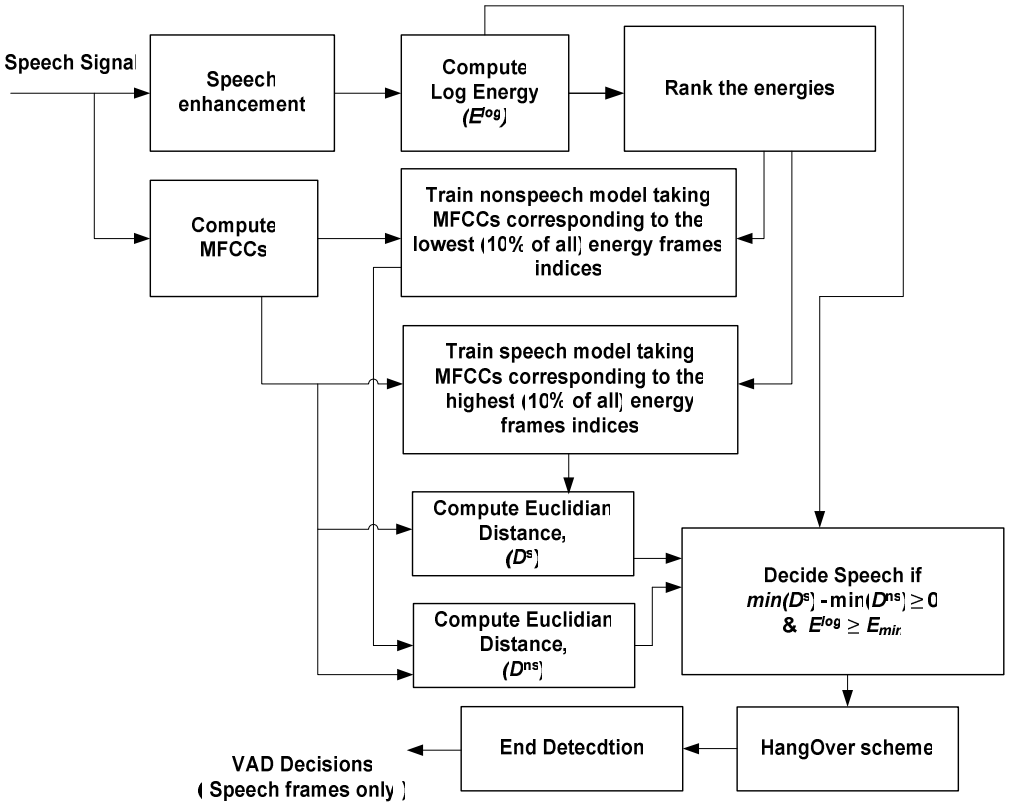

Figure 2: Vector quantization-based self adaptive voice activity detector.

Now, for each frame speech is decided if $\left(\min_k\left(\boldsymbol{D}^s\right) - \min_k\left(\boldsymbol{D}^{ns}\right)\right) \geq 0$ and $\boldsymbol{E}^{\log} \geq \boldsymbol{E}_{\min}$, where $\boldsymbol{E}_{\min} = -75$ dB [9].

Then hangover scheme is used to prevent speech leakage. The hangover scheme does this by reducing the risk of a low-energy portion of speech being falsely classified as non-speech. The final VAD labels are then obtained using an end-point detection algorithm. Robustness of energy-based VAD can be improved by enhancing the noisy speech signal before feeding into the VAD algorithm. Note that in the case of RSR2015 database there is no difference in performances with or without applying speech enhancement. Therefore we did not use speech enhancement in any of the VAD algorithms used in this work

2.3. Sequential GMM (SGMM)-based VAD [10]

A sequential Gaussian mixture model (SGMM)-based VAD algorithm in the Mel-filterbank spectral domain was proposed in [10] that uses an unsupervised learning framework. The input signal is first decomposed into 8-Mel subbands in the frequency domain. Then the log Mel filterbank spectrum is computed and smoothed using a median filter with a window of 5-frames for classification. The Gaussian Mixture Model used comprised two Gaussian distributions, each trying to model either nonspeech or speech. The models were trained using an unsupervised learning process, whereby the initial frames (usually first sixty frames, if the number of frames of an utterance is less than sixty then half of the total frames is taken as initial frames) from a signal were clustered into the two Gaussians, with the distribution with the lowest mean representing nonspeech regions and the distribution with the higher mean representing speech regions. The estimated distributions were also used to determine a decision threshold to discriminate speech from non-speech. Usually, it is chosen as the point between two centers where the probabilities are equal.

Then speech/nonspeech detection is performed in each sub-band, independently of all other subbands, and the results from each subband were used to determine the final output through a voting procedure decided by some threshold determined experimentally. After taking the average of all 8 subbands decisions a decision threshold is computed from it using equation (2) for making speech/nonspeech decisions. A hangover scheme which simply delays the transition from a speech declaration to a non-speech declaration is also implemented to account for the low energy regions of the tail end of utterances. An endpoint detection algorithm is then used to get the final VAD labels.

2.4. Supervised GMM-based VAD [14]

To train speech and nonspeech Gaussian mixture models we select training data from NIST Speaker recognition evaluation (SRE) telephone data from 2004 to 2010 inclusive. we extract 11-dimensional MFCC (including the log energy) features, augmented with their first, second, and third derivative

features making 44-dimensional features, from the all training data. Pre-existing VAD segmentations of reasonable quality are then used to separate speech and nonspeech feature vectors. Two 256-component Gaussian mixture models (GMMs) with diagonal covariance matrices are estimated for the speech and nonspeech models using the separated speech and nonspeech MFCC feature vectors. The GMM estimation process begins with a single Gaussian, which is then iteratively split, mean-perturbed and re-estimated up to 256 components using a Maximum likelihood criterion. Each split doubles the number of Gaussians.

Now, using the trained GMMs, producing a speech/nonspeech VAD segmentation for a new recording (i.e., RSR data in our case) involves the following:

- Extract 44-dimensional MFCC features for all the recordings whose VAD segmentations are needed.
- Compute the log likelihoods, speech log likelihood LL^s and nonspeech log likelihood LL^{ns}, of each feature vector with respect to each trained GMM.
- Apply a median filter of length of 41 to smooth the decision boundaries and then compute the log likelihood ratio (LLR), $LLR = LL^s - LL^{ns}$. The length of LLR vector is same as the number of frames in each feature, i.e., one LLR for each frame
- Now, choose speech if $LLR > \tau$ where $\tau = 0.1$.

2.5. Unsupervised GMM-based VAD [14]

The unsupervised Gaussian mixture model (GMM)-based VAD, shown in fig. 3, is conceptually similar to the VQ-based self adaptive VAD [9] described in section 2.2. In VQ-based VAD speech and nonspeech models are estimated using *k*-means (with $k = 16$) clustering whereas in this case they are trained using 16-component GMMs with diagonal covariance matrices. In unsupervised GMM-based VAD *k*-means clustering is used just for initialization.

In unsupervised GMM-based VAD, producing speech/nonspeech VAD segmentations for an audio recording at hand involves the following steps:

- compute the log energy $E^{\log}$ frame by frame, sort the energies and take the lowest and highest (e.g., 10% of all frames in each case) energy frame indices.
- determine the energy threshold θ from the sorted energies using equation (2).
- compute the MFCC (12-dimensional including the 0th cepstral coefficients, no feature normalization method is applied) features from the observed signal.
- train a 16-components GMM for speech $\lambda^s = \left(\left\{w_c^s\right\}, \left\{\mu_c^s\right\}, \left\{\Sigma_c^s\right\}\right)$ by taking the MFCCs corresponding to the highest energy frame indices. Similarly, by taking MFCCs that corresponds to the lowest energy frame indices train a 16-components GMM for nonspeech $\lambda^{ns} = \left(\left\{w_c^{ns}\right\}, \left\{\mu_c^{ns}\right\}, \left\{\Sigma_c^{ns}\right\}\right)$ where $c = 1, 2, ..., C$ is mixture component index and C represents number of mixture components.
- compute speech log likelihood LL_s of each feature with respect to the trained speech model λ^s. Similarly, given trained nonspeech model λ^{ns} compute nonspeech log likelihood LL_{ns}.
- Compute the log likelihood ratio LLR by simply subtracting nonspeech log likelihood from the speech log likelihood. Smooth the LLR using a moving averaging filter with a sliding window of 23-frames. Determine a threshold θ_{llr} from the sorted likelihood ratio using equation (2).
- Choose speech if $LLR \geq \theta_{llr}$ and $E^{\log} \geq \theta$ otherwise nonspeech.
- Then hangover scheme is used to prevent speech leakage. The hangover scheme does this by reducing the risk of a low-energy portion of speech being falsely classified as non-speech. The final VAD labels (contains only speech frames) are then obtained using an end-point detection algorithm.

The RSR2015 corpus was collected in office environment using 6 portable devices, i.e., there are different channel distortions. For noisy corpus under additive and reverberant environments robustness of this VAD can be improved by simply enhancing the signal using a spectral subtraction technique before feeding into this VAD or by incorporating following changes:

1. Since energy is not a robust feature, specifically in low signal-to-noise condition, its robustness can be improved using spectral subtraction as a pre-processor [4].
2. Instead of MFCCs, robust features such as one proposed in [15], can be used for estimating the GMMs.

3. Experiments and results

Speaker recognition experiments are carried out on the female trials of the RSR2015 corpus. Following six VAD algorithms (described in section 2) are used for performance evaluation: ***Energy-based VAD*** [13], ***Energy-based VAD I, VQ-based VAD*** [9], ***sequential GMM(SGMM)-based VAD*** [10], ***GMM-based VAD (supervised)*** [6, 14], and ***GMM-based VAD (unsupervised).***

Performance evaluation metrics used in this work are : the Equal Error Rate (EER), the old normalized minimum detection cost function ($\text{minDCF}_{\text{Old}}$) and the new normalized minimum detection cost function ($\text{minDCF}_{\text{new}}$). $\text{minDCF}_{\text{Old}}$ and $\text{minDCF}_{\text{new}}$ correspond to the evaluation metric for the NIST SRE in 2008 and 2010, respectively.

3.1. Speech Corpus

RSR2015 (Robust Speaker Recognition 2015), a new speech corpus for text-dependent robust speaker recognition, contains audio recordings from 298 speakers, 142 female and 156 male in 9 sessions each, with a total of 151 hours of speech. The speakers were selected to be representative of the ethnic distribution of Singaporean population, with age ranging from 17 to 42 [18]. The database was collected in office environment using six portable devices (4 smart phones and 2 tablets) from different manufacturers. Each speaker was recorded using three different devices out of the six. Each of the 9 sessions for a speaker is organized into 3 parts [18]:

part I- 30 sentences from the TIMIT database covering all English phones. The average duration of sentences is 3.2 seconds and total duration is 71 hours.

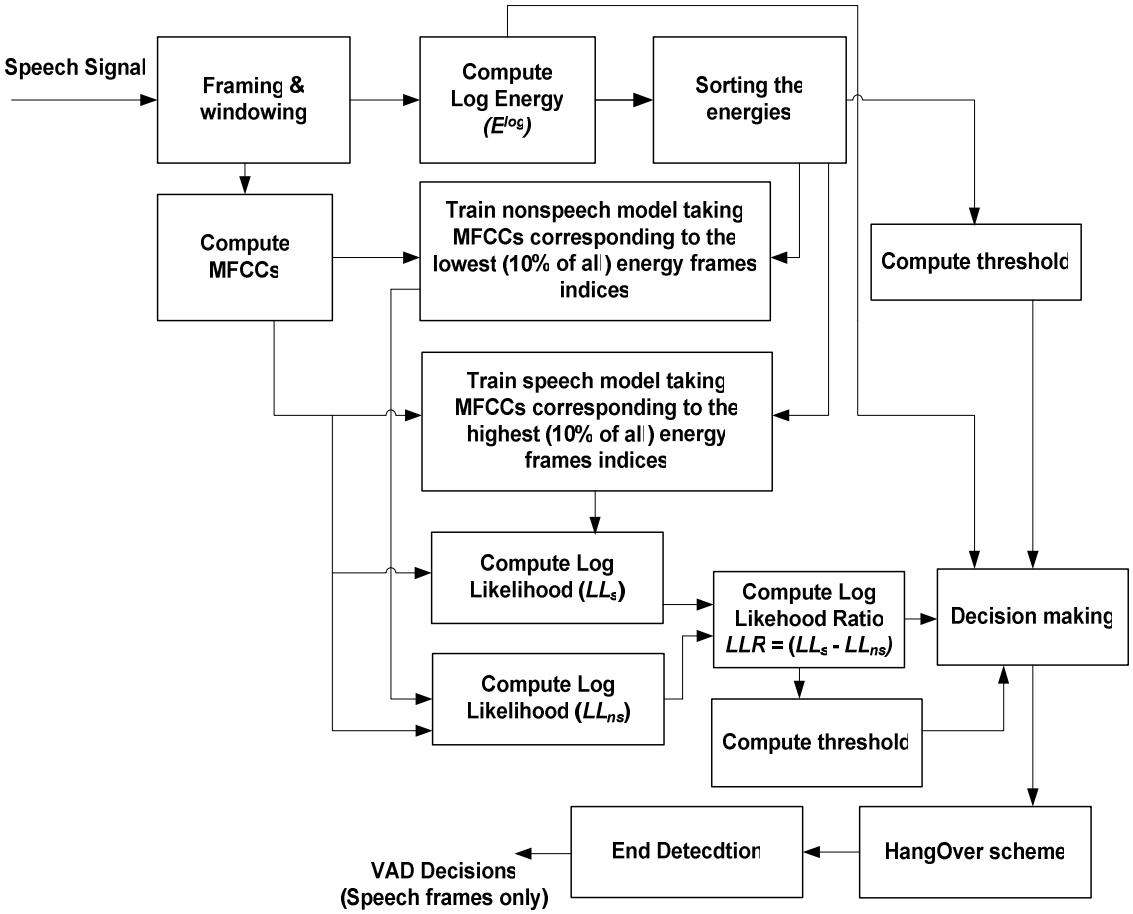

Figure 3: Block diagram of unsupervised Gaussian mixture model (GMM)-based voice activity detector.

part II- 30 short commands designed for the StarHome applications. The average duration of short commands is 2 seconds and total duration is 45 hours.
part III - consists of three 10- and ten 5-session dependent digit strings.
The information about where to get this corpus can be found in [22]. This work deals with a subset of part I of the RSR2015 corpus. Similar to [20] , the background set is used for UBM and Joint Factor Analysis (JFA) training and a restricted test set from the part I evaluation set is used for testing. This restricted test set consists of all the female trails obtained by selecting all the target trials and 50000 high scoring nontarget trials. Working with the restricted test set causes the error rates to increase by a factor of 2 [20, see section 4.1].

3.2. Features Extraction

We extract 20-dimensional MFCC features (including the log energy). First and second derivatives are appended with the static coefficients for a total feature dimension of 60. Then the nonspeech frames are removed using the VAD segmentations of each of the VAD algorithms. Short-term mean and variance normalization (STMVN) with a sliding window of 151-frames is then applied to normalized the features. Features having less than 151 frames are normalized with a full utterance-based MVN

3.3. Experimental Setup

We made six systems for different VAD algorithms considered here. For each system a 512-component gender-independent UBM (universal background model) with diagonal covariance matrices was trained using all background features (63587 recordings with 32770 from 50 male speakers and 30817 from 47 female speakers). Baum-Welch statistics were generated from extracted MFCCs using the trained UBM. Joint Factor Analysis (JFA) was trained using extracted Baum-Welch statistics from all the background data. We use a JFA-based speaker verification framework as proposed in [19, 20] with the rank of eigenchannels matrix = 50 but without UBM adaptation. Please see section 4 of [20] for details about this framework. A restricted test set, as mentioned in section 3.1, is used for evaluation [20].

3.4. Results

In this work we used the part I portion of the RSR2015 corpus. We report results on the female trials of restricted test set of the RSR2015 corpus. The numbers of target and nontarget trials in this restricted test set were 8664 and 50000, respectively. Fig. 4 presents an utterance from the RSR2015 corpus uttered by a female speaker and its VAD segmentations

obtained by all six VAD algorithms considered in this work. It is observed from fig. 4 that compared to with VAD algorithms the unsupervised GMM-based VAD was able to detect speech/nonspeech more accurately. Table 1 presents the equal error rate (EER), $minDCF_{old}$ (minimum normalized detection cost for NIST SRE 2008) and $minDCF_{new}$ (minimum normalized detection cost for NIST SRE 2010) achieved by all six voice activity detection (VAD) algorithms. It is observed from the presented results that both the supervised and unsupervised versions of GMM-based VAD yielded better recognition accuracy than all other VAD algorithms. In terms of all three evaluation metrics unsupervised GMM-based VAD outperformed all other VADs. The relative improvements achieved by unsupervised GMM-based VAD over all other VADs are shown in table 2. Combining energy-based and likelihood ratio-based criteria in VAD algorithm was found helpful. Energy is sensitive to additive noise distortions but its robustness can be improved by incorporating a noise reduction technique as a pre-processor [4, 9].

Table 1:Text dependent speaker verification results in terms of EER, $minDCF_{old}$ and $minDCF_{new}$ on the female trails of restricted test set of the RSR2015 corpus obtained for different VAD algorithms.

	EER (%)	**$minDCF_{old}$**	**$minDCF_{new}$**
Energy-based VAD	2.5	0.096	0.250
Energy-based VAD I	2.3	**0.085**	0.267
VQ-based VAD	2.2	0.087	0.227
GMM-based VAD (supervised)	**2.1**	0.080	0.214
GMM-based VAD (unsupervised)	**2.1**	**0.078**	**0.178**
SGMM-based VAD	2.2	0.083	0.201

Table 2:Percentage relative improvements (RI) obtained by the unsupervised GMM-based voice activity detector (VAD) over all other VADs in EER, $minDCF_{old}$ and $minDCF_{new}$ on the restricted test set. A positive RI indicates reduction in EER, $minDCF_{old}$ and $minDCF_{new}$.

	EER (%)	**$minDCF_{old}$ (%)**	**$minDCF_{new}$ (%)**
Energy-based VAD	16	18.7	28.8
Energy-based VAD I	8.6	8.2	33.3
VQ-based VAD	4.5	10.3	21.5
GMM-based VAD (supervised)	0	2.5	16.8
SGMM-based VAD	4.5	6.0	11.4

4. Conclusions

In this paper we compared several unsupervised and supervised voice activity detection (VAD) algorithms in terms of speaker verification performances on the RSR corpus. It is observed that, if implemented properly, unsupervised VAD can provide similar/better performance than the supervised VAD. Combining energy-based and likelihood ratio-based VAD criterion provided better discrimination of speech from nonspeech. Among all the VAD both the supervised and unsupervised GMM VAD showed better performance in terms of speaker recognition accuracy. The unsupervised GMM-based VAD outperformed all other VADs when speaker recognition performances are compared in terms of all three evaluation metrics, i.e., EER, $minDCF_{old}$ and $minDCF_{new}$.

Our future works are:

- ✓ To evaluate the performance of all the VAD algorithms in different additive and convolutive noise environments.
- ✓ To do fusion of the decisions of different VAD algorithms.
- ✓ To incorporate robust features such as, robust cepstral coefficients [15], long-term signal variability (LTSV) [16], multiband LTSV [17].
- ✓ To evaluate the performances in text-independent speaker recognition task on the NIST SRE 2012 corpora.

5. References

[1] J. Ramirez, J. M. Gorriz, J. C. Segura, Voice Activity detection. Fundamentals and Speech Recognition System Robustness, In M. Grimm and Kroschel, *Robust speech recognition and Understanding*, pp. 1-22, 2007.

[2] 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, Jan. 1999.

[3] 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.

[4] 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.

[5] Lee Ngee Tan, B. J. Borgstrom, Abeer Alwan, "Voice Activity detection using Harmonic Frequency Components in Likelihood ratio Test," Proc. ICASSP, 2010.

[6] ABC System description for NIST SRE 2010. online: http://www.fit.vutbr.cz/research/groups/speech/publi/2010/ABC_system_description%20for%20_NIST%20SRE%202010.pdf

[7] The CRIM System for the 2010 NIST Speaker Recognition Evaluation. online: http://www.crim.ca/perso/patrick.kenny/

[8] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[9] T. Kinnunen, P. Rajan, ”A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data”, *Proc. of* ICASSP, pp. 7229-7233, Vancouver, Canada, May 2013.

[10] Dongwen Ying, Yonghong Yan, Jianwu Dang, Frank K Soong, "Voice Activity Detection Based on an Unsupervised Learning Framework," IEEE Trans. on ASLP, vol. 19, no. 8, November 2011.

[11] F. G. German, D. L. Sun, G. J. Mysore, "Speaker and Noise Independent Voice Activity Detection," Proc. Interspeech, pp. 732-736, August 2013.

[12] K. Yamamoto, F. Jabloun, K. Reinhard, A. Kawamura, "Robust Endpoint detection for Speech recognition based on Discriminative Feature Extraction," Proc. ICASSP, pp. 805-808, 2006.

[13] Endpoint detector, http://www.isip.piconepress.com/projects/speech/software/legacy/signal_detector/index.html.

[14] Luciana Ferrer, Yun lei, Mitchell McLaren, Nicolas Scheffer, Martin Graciarena, Vikramjit Mitra, SRI 2012 NIST Speaker recognition Evaluation System Description, 2012.

[15] M. J. Alam, P. Kenny, D. O'Shaughnessy, "Robust Feature Extraction for Speech Recognition by Enhancing Auditory Spectrum," Proc. INTERSPEECH, Portland Oregon September 2012.

[16] Ghosh P., Tsiartas A., and Narayanan S., “Robust Voice Activity Detection Using Long-Term Signal Variability,” *IEEE Transactions Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600–613, 2011.

[17] Tsiartas A. et. al, "Multi-band long-term signal variability features for robust voice activity detection," Proc. INTERSPEECH, 2013.

[18] A. Larcher, K. A. Lee, B. Ma and H. Li, "RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases", Proc. Interspeech, 2012.

[19] Patrick Kenny, Themos Stafylakis, Pierre Ouellet, Jahangir Alam, " JFA-Based Front Ends for Speaker Recognition," Proc. ICASSP, Florence, Italy, 2014. online: http://www.crim.ca/perso/patrick.kenny/kenny_icassp14.pdf

[20] P. Kenny, T. Stafylakis, J. Alam, P. Ouellet, M. Kockmann, "Joint Factor Analysis for Text-Dependent Speaker Verification," Proc. Odyssey 2014. online: http://www.crim.ca/perso/patrick.kenny/kenny_odyssey14.pdf

[21] 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.

[22] A. Larcher, and H. Li, "The RSR2015 Speech Corpus", SLTC Newsletter, 2012. http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2012-05/the-rss2015-speech-corpus/

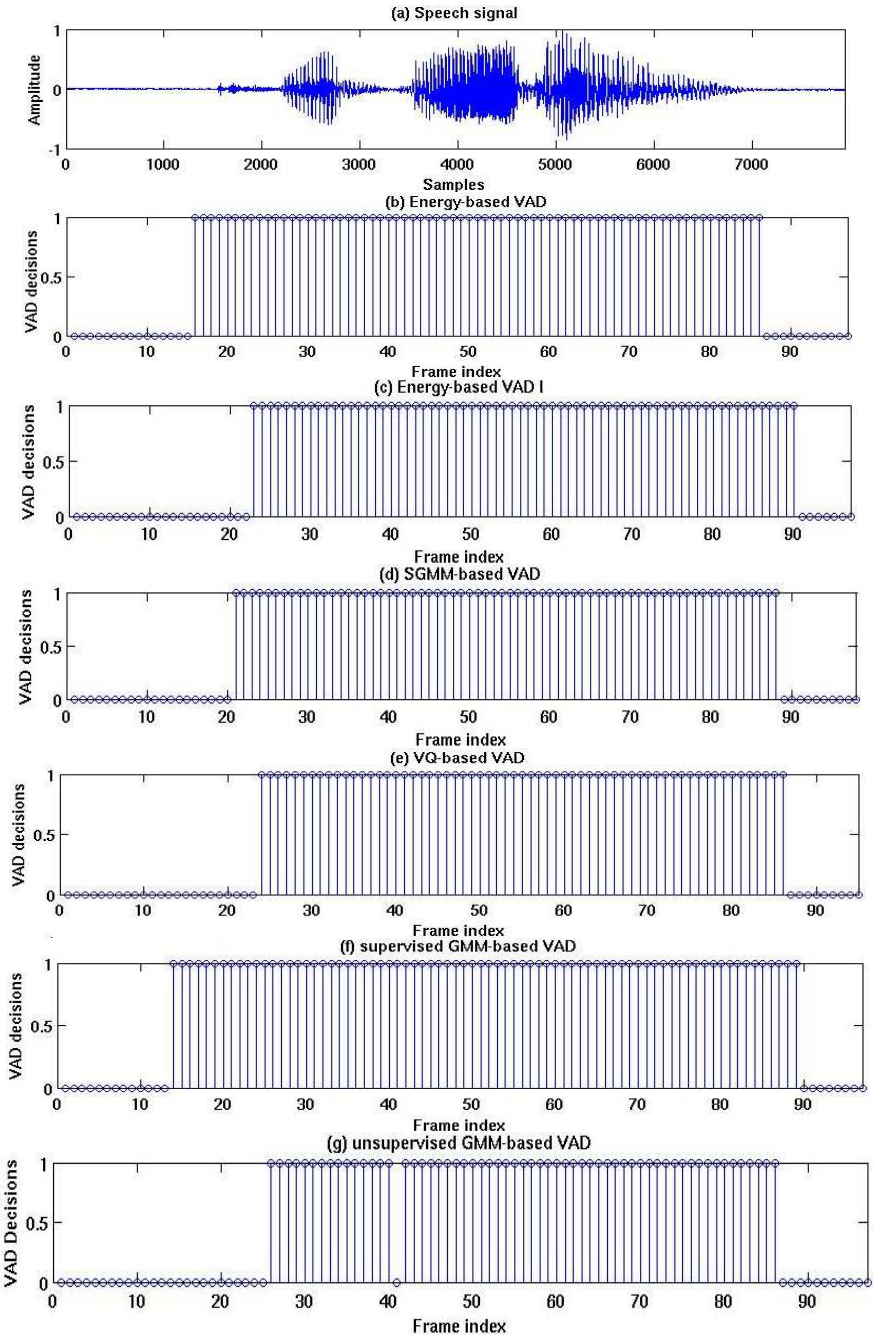

Figure 4: (a) An utterance from the RSR2015 corpus (female)in time domain and its VAD segmentations achieved by (b) Energy-based VAD, (c) energy-based VAD I, (d) sequential GMM(SGMM)-based VAD, (e) VQ-based VAD, (f) supervised GMM-based VAD, and (g) unsupervised GMM-based VAD algorithms.