

Evaluation of Speech Activity for Interview in NIST Speaker Recognition

Nirmal Kumar P.^{1}, Venkatesh C.²*

¹*Department of Electronic and Communication Engineering, Arulmurugan College of Engineering, Thennilai, Karur, Tamil Nadu, India*

²*Professor & Head Department of Computer science and Engineering, Arulmurugan College of Engineering, Thennilai, Karur, Tamil Nadu, India*

**Corresponding Author*

E-Mail Id: nirmalkumar.jan22@gmail.com

ABSTRACT

Interview speech in ongoing NIST Speaker Recognition Evaluations (SREs) has required the improvement of Speech activity detectors (VADs) that can work under extremely low signal to-noise proportion. This paper features the qualities of Interview speech records in NIST SREs and examines the challenges of distinguishing speech/non-speech fragments in these documents. To mitigate these challenges, this paper proposes a VAD that utilizes noise reduction as a pre-preparing step. A methodology to dodge the undesirable impacts of impulsive signals and sinusoidal foundation signals on the VAD is additionally proposed. The proposed VAD is contrasted and the VAD in the ETSI-AMR speech coder for evacuating quietness areas of interview speech documents. The outcomes show that the proposed VAD is more powerful in detecting speech segments under low SNR, prompting a huge exhibition gain in Common Conditions 1–4 of NIST 2008 SRE.

Keywords: *Speech activity detection, far-field microphone, speaker verification, noise reduction, spectral subtraction, NIST speaker recognition evaluations*

INTRODUCTION

Speaker Verification

Speaker verification [2,3] is to authenticate the identity of an individual based on his or her own voice. It is an important branch of biometrics [4,5] and has potential applications in security, access control, password reset, self-service telephone banking, and offender management programmes [6]. For example, Banco Bradesco, a Brazil's private bank, uses Nuance's speaker verification solution to verify its 15 million customers over the phone [7]. In another model, ABN AMRO utilizes Voice Vault's speaker confirmation system in its phone banking administrations [8]. More recently, NAP Personal Banking in Australia and T-mobile of Deutsche Telekom in Netherlands provide voice authentication for their customers [9,10].

The procedure of speaker confirmation can be isolated into two phases: enrollment and verification. During the enlistment stage, a customer speaker is solicited to absolute a set of phrases or sentences. The collected speech is then used to create a client-speaker model corresponding to that speaker.

In a regular confirmation meeting, a speaker guarantees his/her personality; at that point the system prompts the speaker to absolute a particular expression or sentence and looks at the articulation against an objective speaker model relating to the asserted character to settle on a choice. In addition to this prompt-and-response scenario, the conversations of a speaker in telephone calls, meetings, interviews can also be used for enrollment and verification. The last situation has been

utilized in ongoing NIST speaker recognition evaluations (SREs) [11].

Importance of VAD Algorithm in Speaker Verification

NIST SREs [11] have been concentrating on text-free speaker confirmation over phone channels since 1996. In recent years, NIST introduces interview speech into the evaluations. For instance, the speech records in NIST 2008 SRE contain discussion portions of roughly five minutes for phone speech and three minutes for interview speech. In every speech record, about half of the discussion contains speech, the other half being delays or quietness intervals. The incorporation of non-speech intervals in the speech records requires voice activity detection (VAD) in light of the fact that these stretches don't contain any speaker data.

Existing VAD

VAD is an essential part of speech processing and communication systems. Specifically, it helps upgrade system limit and lessen power utilization of handy communication gadgets by means of irregular transmission of coded speech. Early techniques for VAD extract parameter, for example, LPC distance [12], vitality levels, and zero crossing rates [13] from speech signals and contrast this parameter and a lot of thresholds for recognizing the speech regions of an utterance. The threshold is evaluated from non-speech regions of utterances. The detection accuracy of these earlier methods, however, could degrade dramatically under adverse acoustic conditions.

Advanced speech coders normally use extra sophisticated techniques in their VAD. For instance, in Option 1 of ETSI adaptive multi-rate (AMR) coder [1], the decision logic of speech/non-speech is based on a mixture of acoustic information, including pitch, tone, complex-signal correlation, and the energy levels of 9 frequency bands. In Option 2 of the AMR coder, VAD choices rely upon the energy of 16 channels (frequency bands), background noise, channel SNR, frame SNR, and long term SNR. One benefit of this coder is that the VAD choice limit is adjusted powerfully as indicated by the acoustic condition, permitting on-line speech/non-speech discovery under non-fixed acoustic situations. More in recent times, research has concentrated on statistical based VAD where specific frequency receptacles of speech are expected to follow a parametric thickness work [14]. In this methodology, VAD choices depend on a likelihood-ratio test where the mathematical mean of the log-probability proportions of individual frequency containers are assessed from watched speech signals.

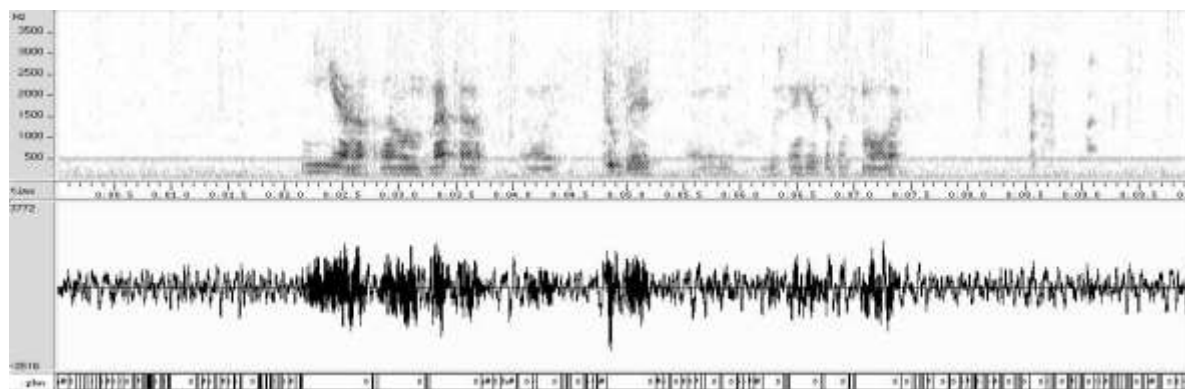
The statistical models can be Gaussian [14]. However, to handle a wide variety of noise conditions, it has been found [15] recently that Laplacian and Gamma models are more appropriate. The kinds of models can likewise be chosen adaptively for various different noise types and SNRs as indicated by an online Kolmogorov-Smirnov test [15]. To improve the heartiness of VAD under unfriendly acoustic condition, relevant data got from different perceptions has been joined into the likelihood ratio tests [16].



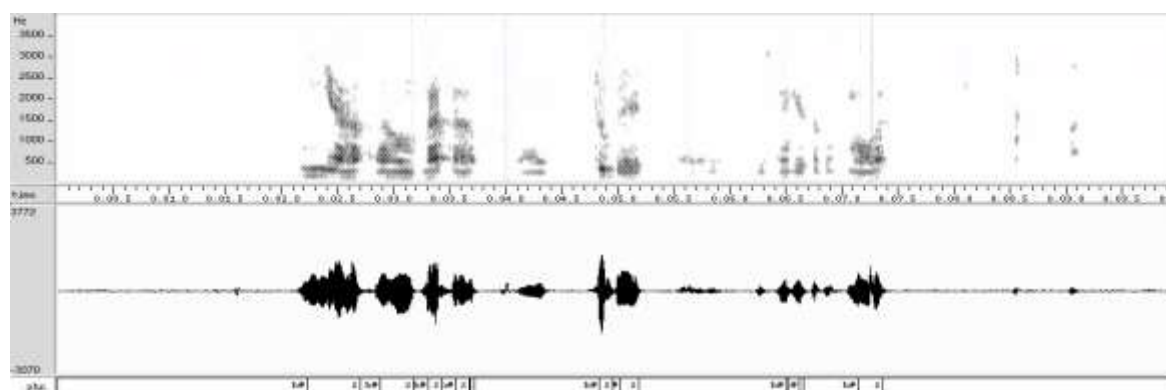
The whole speech file (without denoising)



The whole speech file (with denoising)



A short segment (without denoising)



A short segment (with denoising)



Short Segment

Fig. 1: Spectrogram, Waveform, and Speech/Non-Speech Location of an Interview-Speech Document in NIST 2008 SRE without [(a) and (c)] and with [(b) and (d)] Denoising. (e) VAD Aftereffects of ETSI-AMR Coder, Option 2 [1]. For (c)– (e), the Consequences of VAD are Appeared in the Panels Marked with .phn, with S and h# Representing Speech and Non-Speech Spans, Separately.

Figure 1(e) shows the speech and non-speech segments detected by the ETSI-AMR coder, Option 2. The figure advises that this coder over-estimates the length of speech segments.

In ongoing NIST SREs, a few destinations gave the subtleties of their VAD in the system portrayals. Typically, these systems use energy-based methods that estimate a file-dependent decision threshold according to the maximum energy level of

the file [17]. A few locales utilized the periodicity of speech casings to make speech/non-speech choices [18]. An elective methodology is to utilize the ASR transcript provided by NIST to expel the non-speech segments [19].

Paper Organization

This paper proposes a voice activity detector that is specially design for extracting speech segments from the interview- speech files of NIST SREs. Segment II features the extraordinary qualities of interview speech in ongoing NISR SREs and shows how these attributes cause challenges in extricating the speech segments precisely. Then, Section III argues that spectral subtraction is an essential step in overcoming the difficulties. Further confirmations are then announced in Section IV where the proposed VAD outflanks the VAD in the ETSI AMR coder [1] under the NIST 2008 SRE

INTERVIEW SPEECH IN NIST SRE

The phone speech segments in NIST SREs for the most part have high signal to-noise proportions (SNRs), fundamentally in view of the nearness between the speaker's mouth and the handset in phone speech. The high SNR makes VAD a trivial task. However, for interview speech, different micro- phone types can be used for recording. For instance, twelve microphones have been utilized in recording interview speech in NIST 2008 SRE.

The interview speech records in NIST SREs are unique in that

- Some files have exceptionally low SNR, as demonstrated in Figures 1 and 2;
- Some files comprise low-energy speech superimposed on periodic background signals, as demonstrated in Figure 2; and
- Some files comprise a number of

spikes (impulsive signals) make happen by plosive sounds or the speaker speaking too near to the microphone, as demonstrated in Figure 3.

Depending on the microphone types, some of the interview- speech segments have very low SNRs, causing problems in conventional VAD. Figure 1(a) shows the waveform of an interview-speech file (ftvhv.sph) in NIST 2008 SRE, and Figure 1(c) highlights a short segment of the same file. Evidently, the SNR is very low. This low SNR will cause various blunders in energy-based VAD, as apparent in the lower panel (labelled with .phn) of Figure1(c).

NOISE REDUCTION FOR VAD

Spectral Subtraction as a Preprocessing Step

The unique qualities of interview speech records in NIST SREs require an unpredictable way to deal with identifying the speech segments. Specifically, on account of the low SNR, noise decrease turns into a fundamental preprocessing step. To this end, we have applied spectral subtraction (SS) with a huge over-subtraction factor to evacuate the background noise however much as could be expected before passing the denoised speech to an energy based VAD. We didn't utilize further developed speech enhancement procedures, (for example, MMSE [20] and LSA-MMSE [21]) on the grounds that our attention isn't on the sound quality of the recreated speech. Rather, our attention is on expanding the sign to-noise ratio in the speech regions while simultaneously limiting the signal amplitude in the non-speech regions. Spectral subtraction can meet this prerequisite well without superfluously entangling the entire system.

Denote $x(n, m)$, $y(n, m)$, and $b(n, m)$ as the clean, noisy, and background signal at frame m , respectively. Additionally denote

their comparing frequency spectrum as $X(\omega, m)$, $Y(\omega, m)$, and $B(\omega, m)$, respectively. To estimate the clean speech

from the observed noisy speech, this paper uses the spectral subtraction [22–24] of the form:

$$\hat{X}(\omega, m) = \begin{cases} \sum |Y(\omega, m)| - \alpha_m |\hat{B}(\omega)| e^{j\phi_y(\omega, m)} & \text{if } |Y(\omega, m)| > (\alpha_m + \beta_m) |\hat{B}(\omega)| \\ \beta_m \hat{B}(\omega) e^{j\phi(\omega, m)} & \text{otherwise,} \end{cases} \quad (1)$$

where

$\phi_y(\omega, m)$ is the phase of $Y(\omega, m)$, $B(\omega)$ is the average spectrum of some non-speech regions, $\alpha_m \geq 1$ is an over-subtraction factor for removing background noise, and $0 < \beta_m \ll 1$ is a spectral floor factor ensuring that the recovered spectra never fall below a preset minimum (spectral floor). The over-subtraction factor aims to reduce the background noise as much as possible

when the signal energy is significantly higher than the background noise. When the SNR is low, the spectral floor factor ensures that a low-level of noise is present in the enhanced signal. This noise help stored use the annoying effect of the musical noise that may otherwise be introduced if the recovered spectrum $\hat{X}(\omega, m)$ is set to zero.



Fig. 2: A Short Segment of Low-Energy Interview Speech in NIST 2008 SRE Containing a High-Energy Spike.

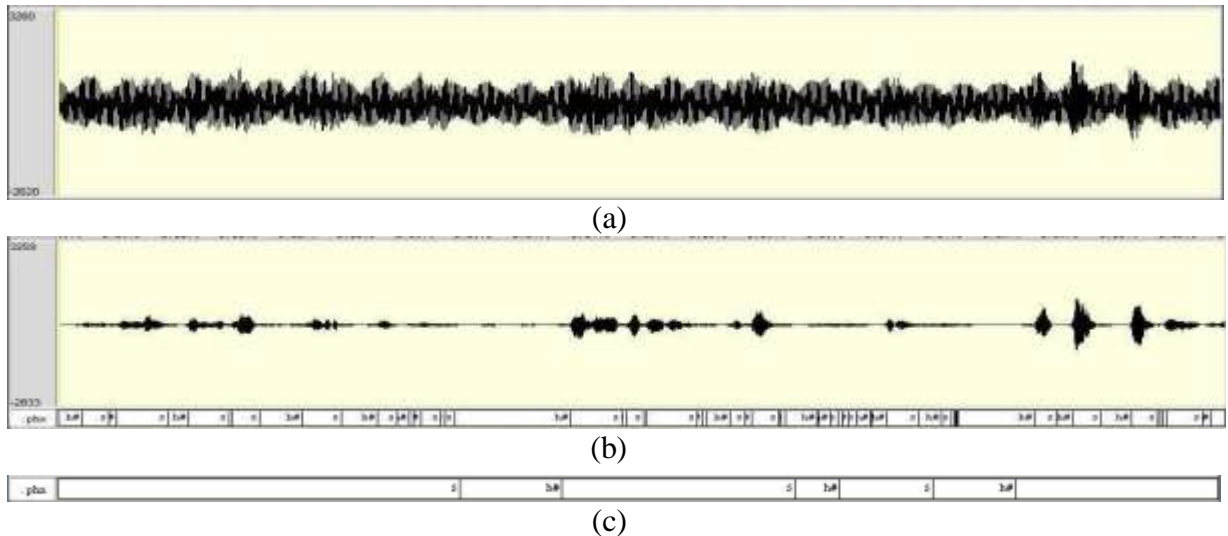


Fig. 3: (a) A Short Segment of Low-Energy Interview Speech in NIST 2008 SRE Superimposed on a Periodic Background. (b) The Same Segment after Spectral Subtraction. The VAD Decisions (S for Speech and h# for Silence) are shown in the Bottom Section. (c) VAD Decisions made by an ETSI-AMRCoder.

The value of α_m and β_m can be computed as follows

$$\alpha_m = -\frac{1}{2}\xi_m + c \quad (\alpha_{\min} \leq \alpha_m \leq \alpha_{\max})$$

$$\beta_m = \begin{cases} \beta_{\min} & \text{if } \xi_m < 1 \\ \beta_{\max} & \text{otherwise} \end{cases}$$

where

$$\xi_m = |Y(\omega, m)|^2 / |\hat{B}(\omega)|^2$$

α_m and β_m is the *a posteriori* SNR, c is a constant ($= 2.5$ in this work), α_{\min} , α_{\max} , β_{\min} , and β_{\max} compel the suitable scope of the over-subtraction factor and the noise floor. These cutoff points are set by the measure of tolerable musical noise in the noise diminished speech. Since melodic commotion isn't a worry in our application (speakers' highlights were separated from the first records rather than the noise diminished files), we set these qualities with the end goal that the speech spectra are over-subtracted, i.e., we expelled however much noise as could reasonably be expected. In this work, we set $\alpha_{\max} = 4$, $\alpha_{\min} = 0.5$, $\beta_{\max} = 0.05$, and $\beta_{\min} =$

0.01.

These qualities were controlled by watching the remade waveform of a several documents.

Figure 4 shows the structure of the proposed VAD, which we allude to as spectral subtraction VAD or basically SS-VAD. Figures 1(b) and 1(d) show a similar speech file and section as in Figures 1(a) and 1(c) however after spectral subtraction. Obviously, with the background Noise to a great extent evacuated, speech and non-speech spans can be accurately recognized by energy based VAD.

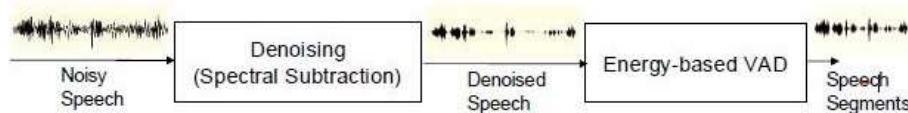


Fig. 4: The Structure of the Proposed VAD for NIST SREs.

To collect more evidences on the advantage of noise removal, we applied (1) energy-based VAD without SS, (2) ETSI AMR, and (3) energy-based VAD with SS to extract the speech segments of 6,249 files in NIST'05-08.

For each record, we utilized the three indicators to separate the speech segments and registered the proportion between speech segment length and complete signal length. The circulations of speech segment length to total signal length proportion are appeared in Figure 5. The figure shows that without noise removal, the identifier

mistakenly decides numerous non-speech segments as speech segments in an enormous number of speech records, as obvious by the high frequency of events at proportion 0.9–1.0. Then again, with noise removal, the finder thinks about that in numerous speech files, half of the all-out signals contain speech. The ETSI AMR lies in the middle of VAD with noise evacuation and VAD without noise removal.

Threshold Determination and VAD Decision Logic

The presence of impulsive signals (spikes)

also causes problems in determining the VAD decision threshold, because the spikes affect the maximum SNR in the file. If the decision threshold is based on the background amplitude and the maximum amplitude, the presence of these spikes will lead to overestimation of the decision threshold, causing low-energy. The nearness of impulsive signals (spikes)

additionally messes up deciding the VAD choice threshold, on the grounds that the spikes influence the most extreme SNR in the record. In the event that the decision threshold depends on the background amplitude and the most extreme amplitude, the nearness of these spikes will prompt overestimation of the choice threshold, causing low-energy.

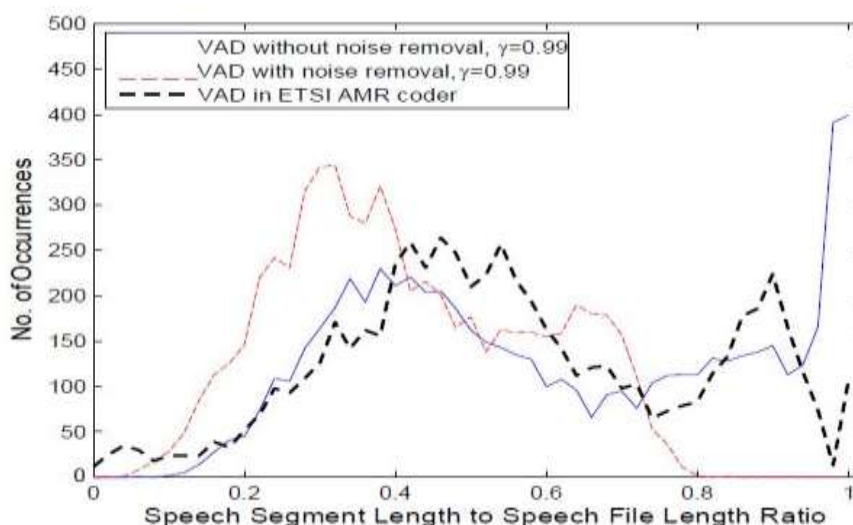


Fig. 5: Distribution of Speech-Segment-Length to Total-Signal-Length Ratio Determined by three VAD Detectors: Energy-Based VAD without Noise Removal (Blue), Energy-Based VAD with Noise Removal (Red Dashed), and VAD (Option 2) in ETSI-AMR Coder (Black Dashed-Dot).

The Graph between Number of Occurrences and Speech Segment Length to Speech File Length Ratio

VAD without noise removal, $\gamma=0.99$

VAD with noise removal, $\gamma=0.99$

VAD in ETSI AMR coder

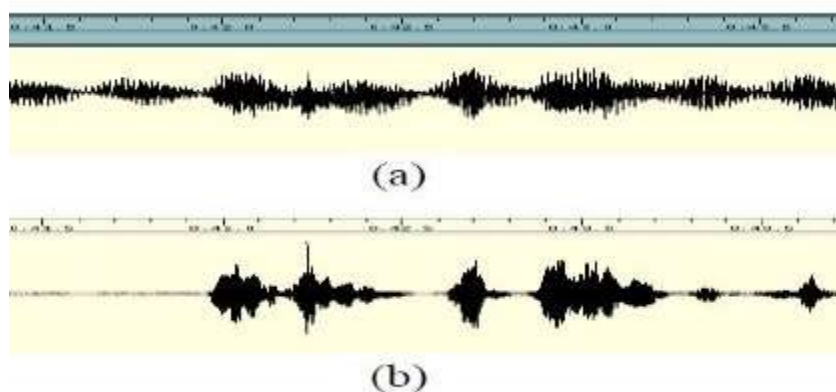


Fig. 6: (a) A Short Segment of Periodic Background in NIST 2008 SRE. (b) The Same Segment after Spectral Subtraction.

Speech segments to be mistakenly detected as non-speech. To address this problem, we have developed a strategy to prevent the spikes from interfering the threshold estimation. More precisely, a fixed percentage (e.g., 10%) of the speech file is expected to comprise signal peaks (including spikes). Then, the smallest magnitude of these peaks is determined. The VAD decision threshold θ is a linear combination of the smallest of the signal peaks and the mean of background amplitude μ_b , as follows:

$$\theta = \gamma \mu_b + (1 - \gamma) \times \min\{ap_1, \dots, ap_L\}, \quad (3)$$

where $0 < \gamma < 1$ is a weighting factor and $\{ap_1, \dots, ap_L\}$ are amplitudes of L frames with the largest amplitude. Note that L can't be excessively huge, in any case the rank rundown may incorporate the peaks of some high-energy speech outlines, which will prompt under-estimation of θ . Be that as it may, when L is excessively little, some medium-amplitude spikes will be remembered fondly. In this work L was set to 1% of the all-out number of frames in the speech file. It was found that the influence of spikes can be largely eliminated by using the minimum amplitude in this ranked list. When the VAD decision threshold has been gotten, speech segment can be recognized by looking at the amplitude of each frame in the record with the threshold. Those casings with amplitude bigger than the edge are considered as speech frame. Be that as it may, some speech file contains portions with a huge DC counterbalance after spectral subtraction, as delineated in Fig. 6. These segments ought to be considered as non-speech. Consequently, another choice rationale is included: Frames with amazingly low zero-crossing rate (littler than 10% of background zero-crossing rate) are considered as non-speech.

EXPERIMENTS AND RESULTS

VAD algorithms are typically assessed by looking at the VAD choices on clean

speech against the VAD choices on noise polluted speech [25]. The closer the choices between the VAD under these two conditions, the more vigorous is the VAD algorithm. However, in NIST SREs, the noisy- speech files do not have their clean counterparts. In this way, there are no references for the VAD choices on noisy speech except if hand labeling is performed. Given the huge number of speech files in NIST SREs, hand labeling is not feasible. A potential arrangement is to utilize the presentation records (e.g., EER, minimum DCF, and DET) of speaker check. This is the methodology embraced in this paper.

Speech Data, Features, and Scoring

NIST 2005–2008 Speaker Recognition Evaluation (SRE)1 were utilized in the analyses. NIST'05 and NIST'06 SRE were utilized as development data, and NIST'08 was utilized for execution evaluations.2 only male speakers in these corpora were utilized.

The core task(short2-short3) of NIST'08 has eight basic conditions. This paper centers around Common Conditions 1 to 4 (CC1–CC4), in light of the fact that these four conditions include interview with speech. For instance, CC3 mirrors the presentation of frameworks that were prepared and tried on various microphones in the interview recordings.

For each utterance, an energy-based VAD, the ETSI-AMR coder, and the proposed SS-VAD were used to remove the silence regions, resulting in three segmentation files for subsequent feature extraction (see below). For the SS-VAD, diverse values of the weighting factor (γ in Eq. 3) were applied to the speech files in NIST'08. For the speech files in NIST'05 and NIST'06 used for creating the UBM and T norm models, the γ weighting factor was set to 0.95.3.

Hereafter, all NIST SREs are abbreviated as NIST'XX, where XX stands for the year of evaluation.

We fixed the weighting factor for all speech files utilized for making the UBM

and Tnorm models since we expect that the ideal estimation of this boundary can be gotten during framework advancement.

Figure 7 shows the pseudo code of the proposed SS-VAD.

$\mathbf{x}, \dots, \mathbf{x}$ - speech and non-speech frames
 α, α - limits of over-subtraction factor (refer to Eq. 2)
 β, β - limits of spectral floor factor (refer to Eq. 2)
 γ - combination weight (refer to Eq. 3)
 $\mathbf{y}, \dots, \mathbf{y}$ - denoised speech and non-speech frames
 $\mathbf{b}, \dots, \mathbf{b}$ - background frames (typically $K = 0.05N$)
 $\mathbf{p}, \dots, \mathbf{p}$ - peak frames (typically $L = 0.01N$)
 a, \dots, a - amplitude of denoised frames
 a, \dots, a - amplitude of background frames
 a, \dots, a - amplitude of peak frames
 z, \dots, z - zero-crossing rate of denoised frames
 z, \dots, z - zero-crossing rate of background frames
 θ - VAD decision threshold
 $\mathbf{s}, \dots, \mathbf{s}$ - speech frames

```
// Denoise input signal using spectral subtraction, refer to Eq. 1
[y, ..., y] = spectral_subtraction([x, ..., x], alpha, alpha, beta, beta);
// Remove DC offset
[y, ..., y] = remove_dc_offset([y, ..., y]);
// Find the background frames by searching for K frames with the lowest amplitude among the N frames in the denoised speech
[b, ..., b] = find_bkg_frames([y, ..., y]);
// Find the peak frames by searching for L frames with the largest amplitude among the N frames in the denoised speech
[p, ..., p] = find_pk_frames([y, ..., y]);
// Determine VAD threshold theta based on the mean of background frames and the minimum amplitude of peak frames
[a, ..., a] = amplitude([b, ..., b]);
[a, ..., a] = amplitude([p, ..., p]);
mu = mean([a, ..., a]); theta = gamma * mu + (1 - gamma) * min([a, ..., a]);
(theta == 0 || theta > 0.2 * mean([a, ..., a]))
theta = 0.2 * mean([a, ..., a]);

// Detect speech frames by comparing the smoothed amplitude of [y, ..., y] with threshold theta
// Consider frames with extremely low zero-crossing rate as non-speech
[a, ..., a] = amplitude([y, ..., y]); [a, ..., a] = moving_average([a, ..., a]);
[z, ..., z] = moving_average([z, ..., z]);
t = 1;
for i = 1, ..., N
    (a > theta && z > 0.1 * mean([z, ..., z]))
        s = x;
        t = t + 1;

// end of SSVAD algorithm
```

Fig. 7: Pseudo Code of the Proposed VAD.

Table I summaries these four common conditions in NIST'08.

Table I: The Training and Test Speech Types Used in Common Conditions 5 1 to 4 in NIST'08 (Male Speakers).

Common Condition	Train/Test Condition	No. of Targets	No. of Trials
1	All Interview speech	622	14405
2	Interview speech, same micro phone type for training and test	125	731
3	Interview speech, different microphone types for training and test	622	13674
4	Interview speech for training, telephone speech for test	622	5048

Twelfth-order MFCCs [26] in addition to their first subordinate were extricated from the speech regions of the articulation, prompting 24-diminish acoustic vectors. Cepstral mean standardization [27] was applied to the MFCCs, trailed by feature warping [28]. We used GMM-SVM [29] as target-speaker models. Specifically, interview utterances from the male speakers of NIST'05 and NIST'06 were used for creating a 512-center, gender-dependent universal background model (UBM). MAP adaptation [30], with relevance factor set to 16, was then performed for each of the target-speakers to create target-dependent GMMs. The similar MAP adaptation was too applied to 300 background speakers (similarly from NIST'05 and '06) to create 300 impostor GMMs. The mean vectors of these GMMs were stacked to form 12288-dim GMM-super vectors [29]. For each target speaker, his target-dependent GMM-supervector and the background GMM-supervectors were used to train a GMM-SVM speaker model.

To decrease channel effects, 81 male speakers from NIST'05 and NIST'06 were utilizes for assessing the gender-dependent NAP matrices [31]. Each of these speakers has at least 8 utterances. The NAP corank was set to 128 for both genders. Three hundred male utterances from NIST'05 were utilized for making T-norm speaker models [32]. The same set of background speakers used for creating the target-speaker SVMs were used for creating the T-norm SVMs.

RESULTS AND DISCUSSIONS

Table II shows the equal error rate (EER) and minimum decision cost (min DCF) accomplished by the three VAD strategies. The outcomes unequivocally propose that preprocessing the noisy sound files by spectral subtraction is a promising thought. With SS, the VAD lessens the EER by 21% in CC1. The outcomes likewise propose that the best scope of γ in Eq. 3 is somewhere in the range of 0.95 and 0.99. When this worth dips under 0.95, the performance degrades quickly. 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. Be that as it may, the threshold can't thoroughly depends on the background amplitude, as the EER and minDCF increment when γ increments from 0.99 to 1.0.

Figure 8 shows the DET execution (under CC1) of the three VAD techniques. The outcomes show that SS-VAD accomplished a noteworthy lower error rates than the ETSI-AMR coder for a wide scope of operating points. Note that the performance of all systems in Table 2 under CC4 is poor. Note that the exhibition of all frameworks in Table 2 under CC4 is poor. This is on the grounds that the NAP matrix was prepared on interview speech just, *i.e.*, the matrix isn't improved for the condition where interview speech is utilized for preparing and phone speech is utilized for testing.

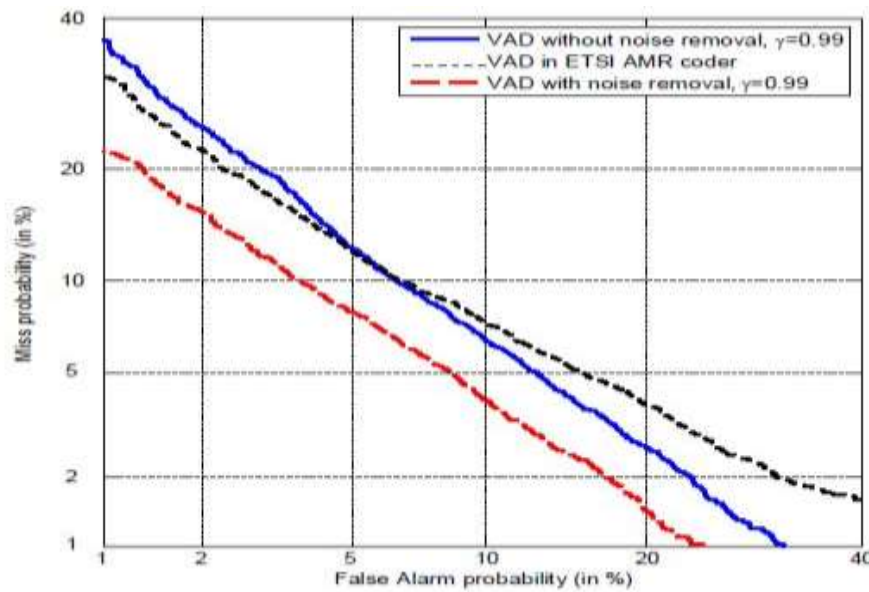


Fig. 8: DET Performance on Common Condition 1 in NIST'08 (Male).

Further work is required to create a NAP matrix to deal with this situation. We additionally notice that the VAD in AMR over-gauges the length of speech segments in light of the fact that the VAD is streamlined for speech coding. One potential arrangement is to build the estimation of the channel-noise smoothing-factor in the coder (on in [1]) with the goal that the VAD turns out to be more rigid.

CONCLUSIONS

A voice activity detector specially designed for extracting speech segments from the interview-speech files in

NISTSREs was proposed and evaluated under the NIST 2008 SRE proto- col. A few ends can be drawn from the investigations accomplished in this work: (1) noise decrease is of essential significance for VAD under amazingly low SNR, (2) it is imperative to evacuate the sinusoidal background found in NIST SRE sound files as this sort of background signal could prompt numerous false recognition in energy based VAD, and (3) our proposed spectral subtraction VAD beats the VAD in an advanced speech coder (ETSI-AMR, Option 2) in speaker confirmation.

Table 2: Performance on NIST 2008 SRE under Common Conditions (CC) 1 to 4. Γ in the 2nd Column is the Weighting Factor in eq. 3 for the Interview-Speech Files in NIST'08. Baseline: Energy-Based VAD without Noise Removal. ETSI-AMR: VAD in AMR Coder. SS-VAD: the Proposed Spectral-Subtraction VAD.

VAD Method	γ	EER (%)				Minimum DCF			
		CC1	CC2	CC3	CC4	CC1	CC2	CC3	CC4
Baseline	0.95	8.28	1.93	8.08	13.61	0.0412	0.0085	0.0406	0.0529
Baseline	0.99	8.14	3.57	7.71	12.68	0.0433	0.0153	0.0411	0.0480
ETSI-AMR	–	8.51	1.53	8.54	11.05	0.0400	0.0032	0.0396	0.0424
SS-VAD	0.00	16.69	10.21	16.72	19.49	0.0685	0.0524	0.0657	0.0742
SS-VAD	0.80	13.29	5.51	13.40	16.36	0.0550	0.0121	0.0549	0.0542
SS-VAD	0.90	9.09	1.93	9.25	12.21	0.0406	0.0089	0.0408	0.0394
SS-VAD	0.95	6.94	1.12	7.08	9.99	0.0347	0.0081	0.0349	0.0380
SS-VAD	0.99	6.44	1.12	6.37	9.64	0.0319	0.0065	0.0319	0.0376
SS-VAD	1.00	8.94	1.83	8.82	13.11	0.0418	0.0085	0.0410	0.0527

REFERENCES

1. ETSI, E. 301 708 V7. 1.1 (1999-12), Digital cellular telecommunications system (Phase 2+): Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels. *General description(GSM 06.94 version 7.1. 1 Release 1998)*.
2. Bimbot, F., Bonastre, J. F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., & Reynolds, D. A. (2004). A tutorial on text-independent speaker verification. *EURASIP Journal on Advances in Signal Processing*, 2004(4), 101962.
3. Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, 52(1), 12-40.
4. Jain, A. K., Flynn, P., & Ross, A. A. (Eds.). (2007). *Handbook of biometrics*. Springer Science & Business Media.
5. Kung, S. Y., Mak, M. W., & Lin, S. H. (2005). *Biometric authentication: a machine learning approach* (pp. 27-49). New York: Prentice Hall Professional Technical Reference.
6. (2009). Speaking up for biometrics. *Biometric Technology Today*, 8,9-11.
7. (2005). Financial success for biometrics? *Biometric Technology Today*, 13(4),9-11.
8. (2006). ABN AMRO to roll out speaker verification next term system for telephone banking. *Biometric Technology Today*, 14(7-8), 3-4.
9. (2009). Speaker verification finds its voice in Australia. *Biometric Technology Today*,17(6),4.
10. (2009). T-mobile trials speaker verification. *Biometric Technology Today*, 11, 2-3.
11. <http://www.itl.nist.gov/iad/mig//tests/sre/>
12. Rabiner, L., & Sambur, M. (1977, May). Voiced-unvoiced-silence detection using the Itakura LPC distance measure. In *ICASSP'77. IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 2, pp. 323-326). IEEE.
13. Junqua, J. C., Reaves, B., & Mak, B. (1991). A study of endpoint detection algorithms in adverse conditions: incidence on a DTW and HMM recognizer. In *Second European conference on speech communication and technology*.
14. Sohn, J., Kim, N. S. & Sung W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1), 1-3.
15. 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.
16. Ramírez, J., Segura, J. C., Górriz, J. M., & García, L. (2007). Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8), 2177-2189.
17. Kinnunen, T., Saastamoinen, J., Hautamäki, V., Vinni, M., & Franti, P. (2009, April). Comparing maximum a posteriori vector quantization and gaussian mixture models in speaker verification. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4229-4232). IEEE.
18. Hautamäki, V., Tuononen, M., Niemi-Laitinen, T., & Fränti, P. (2007, October). Improving speaker verification by periodicity based voice activity detection. In *Proc. 12th Int. Conf. Speech and Computer (SPECOM'2007)* (Vol. 2, pp. 645-650).
19. Dalmaso, E., Castaldo, F., Laface, P., Colibro, D., & Vair, C. (2009, April). Loquendo-Politecnico di Torino's

- 2008 NIST speaker recognition evaluation system. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4213-4216). IEEE.
20. 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.
21. 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(2), 443-445.
22. Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, 27(2), 113-120.
23. Deller, J. R. (1993). JR., JG Proakis, and JHL Hansen. *Discrete-time Processing of speech signals*, 179, 180.
24. 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.
25. Basbug, F., Nandkumar, S., & Swaminathan, K. (1999, June). Robust voice activity detection for DTX operation of speech coders. In *1999 IEEE Workshop on Speech Coding Proceedings. Model, Coders, and Error Criteria (Cat. No. 99EX351)* (pp. 58-60). IEEE.
26. Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4), 357-366.
27. Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America*, 55(6), 1304-1312.
28. Pelecanos, J., & Sridharan, S. (2001). Feature warping for robust speaker verification. In *Proc. Speaker Odyssey*, 213-218.
29. Campbell, W. M., Sturim, D. E., & Reynolds, D. A. (2006). Support vector machines using GMM super vectors for speaker verification. *IEEE signal processing letters*, 13(5), 308-311.
30. 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.
31. Campbell, W. M., Sturim, D. E., Reynolds, D. A., & Solomonoff, A. (2006, May). SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In *2006 IEEE International conference on acoustics speech and signal processing proceedings* (Vol. 1, pp. I-I). IEEE.
32. Auckenthaler, R., Carey, M., & Lloyd-Thomas, H. (2000). Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1-3), 42-54.