rVAD

An unsupervised segment-based robust voice activity detection method

Tan, Zheng Hua; Sarkar, Achintya kr; Dehak, Najim

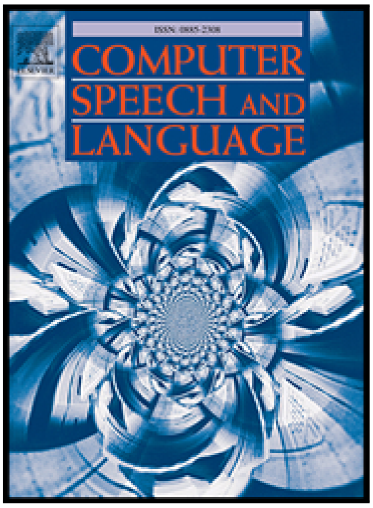
Published in:
Computer Speech and Language

DOI (link to publication from Publisher):
10.1016/j.csl.2019.06.005

Publication date:
2020

Document Version
Accepted author manuscript, peer reviewed version

Citation for published version (APA):
Tan, Z. H., Sarkar, A. K., & Dehak, N. (2020). rVAD: An unsupervised segment-based robust voice activity detection method. *Computer Speech and Language*, *59*, 1-21. https://doi.org/10.1016/j.csl.2019.06.005



rVAD: An Unsupervised Segment-Based Robust Voice Activity Detection Method

Zheng-Hua Tan, Achintya kr. Sarkar, Najim Dehak

PII: S0885-2308(19)30092-0
DOI: https://doi.org/10.1016/j.csl.2019.06.005
Reference: YCSLA 1014

To appear in: *Computer Speech & Language*

Received date: 20 March 2019
Accepted date: 3 June 2019

Please cite this article as: Zheng-Hua Tan, Achintya kr. Sarkar, Najim Dehak, rVAD: An Unsupervised Segment-Based Robust Voice Activity Detection Method, *Computer Speech & Language* (2019), doi: https://doi.org/10.1016/j.csl.2019.06.005

Highlights

- Proposed an unsupervised segment-based method for robust voice activity detection.
- Proposed modified rVAD that uses computationally fast spectral flatness calculation.
- Evaluated rVAD in terms of VAD performance using RATS and Aurora-2 databases.
- Evaluated rVAD in terms of speaker verification performance using RedDots 2016.
- rVAD showed favorable performance on various difficult tasks over existing methods.

rVAD: An Unsupervised Segment-Based Robust Voice Activity Detection Method

Zheng-Hua Tan[a]*, Achintya kr. Sarkar[a,b], Najim Dehak[c]

[a]*Department of Electronic Systems, Aalborg University, Denmark*
[b]*School of Electronics Engineering, VIT-AP University, India*
[c]*Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA*

Abstract

This paper presents an unsupervised segment-based method for robust voice activity detection (rVAD). The method consists of two passes of denoising followed by a voice activity detection (VAD) stage. In the first pass, high-energy segments in a speech signal are detected by using *a posteriori* signal-to-noise ratio (SNR) weighted energy difference and if no pitch is detected within a segment, the segment is considered as a high-energy noise segment and set to zero. In the second pass, the speech signal is denoised by a speech enhancement method, for which several methods are explored. Next, neighbouring frames with pitch are grouped together to form pitch segments, and based on speech statistics, the pitch segments are further extended from both ends in order to include both voiced and unvoiced sounds and likely non-speech parts as well. In the end, *a posteriori* SNR weighted energy difference is applied to the extended pitch segments of the denoised speech signal for detecting voice activity. We evaluate the VAD performance of the proposed method using two databases, RATS and Aurora-2, which contain a large variety of noise conditions. The rVAD method is further evaluated, in terms of speaker verification performance, on the RedDots 2016 challenge database and its noise-corrupted versions. Experiment results show that rVAD is compared favourably with a number of existing methods. In addition, we present a modified version of rVAD where computationally intensive pitch extraction is replaced by computationally efficient spectral flatness calculation. The modified version significantly reduces the computational complexity at the cost of moderately inferior VAD performance, which is an advantage when processing a large amount of data and running on low resource devices. The source code of rVAD is made publicly available.

Keywords: *a posteriori* SNR; energy; pitch detection; spectral flatness; speech enhancement; voice activity detection; speaker verification

1. Introduction

Voice activity detection (VAD), also called speech activity detection (SAD), is widely used in real-world speech systems for improving robustness against additive noises or discarding the non-speech part of a signal to reduce the computational cost of downstream processing [1]. It attempts to detect the presence or absence of speech in a segment of an acoustic signal. The detected non-speech segments can subsequently be abandoned to improve the overall performance of these systems. For instance, cutting out noise only segments can reduce the error rates of speech recognition and speaker recognition systems [2, 3, 4].

VAD methods can be broadly categorized into supervised and unsupervised methods. Supervised methods formulate VAD as a classical classification problem and solve it either by training a classifier directly [5], or by training statistical models for speech and non-speech separately and then making VAD decisions based on a log-likelihood ratio test (LRT) [6]. These methods require labelled speech data and their performances

*Corresponding author. Email address: zt@es.aau.dk. This work was done in part while the author was visiting Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge MA, USA.

are highly dependent on the quality of the labelled data, in particular how well training and test data match each other. Supervised methods are often able to outperform unsupervised methods under matched conditions, but they can potentially break down under mismatched conditions. In [7], for example, a deep learning method, experimented on the robust automatic transcription of speech (RATS) database [8], demonstrates a very good VAD performance for seen environments, but the gap between unseen and seen environments is very significant, by almost an order of magnitude difference in terms of detection cost function (DCF), but what matters most is the performance on unseen environments. Unsupervised VAD methods include metrics methods and model based ones. Metrics methods rely on the continuous observation of a specific metric, such as energy and zero-crossing rate, followed by a simple threshold-based decision stage [9]. On the other hand, model based methods build separate models for speech and non-speech, followed by an LRT test together with threshold-based decision for statistical models [10] (similarly to the settings of supervised methods but without using labelled data) or a distance comparison for non-statistical models [11] to classify a segment as speech or non-speech. Semi-supervised VAD has also been studied, e.g., semi-supervised Gaussian mixture models (SSGMM) based VAD for speaker verification in [12].

VAD is a binary classification problem involving both feature extraction and classification. Various speech features can be found in literature such as energy, zero-crossing rate, harmonicity [13], perceptual spectral flux [14], Mel-frequency cepstral coefficient (MFCC) [11], power-normalized cepstral coefficients (PNCCs) [15], entropy [16], Mel-filter bank (MFB) outputs [17] and *a posteriori* signal-to-noise ratio (SNR) weighted energy distance [3, 18]. For the purpose of modelling and classification, popular techniques include Gaussian models [19], Gaussian mixture models (GMM) [2, 14, 20], super-Gaussian models and their convex combination [9], i-vector [15, 21], decision trees [22], support vector machines [23], and neural network models (including deep models) [5, 7, 24].

Although much progress has been made in the area of VAD, developing VAD methods that are accurate in both clean and noisy environments and are able to generalize well under unseen environments is still an unsolved problem. For example, the ETSI advanced front-end (AFE) VAD [25], a commonly referred VAD method, performs very well in noisy environments, but poorly in noise-free conditions, and it is primarily suitable for dealing with stationary noise. The MFB output based VAD [17] and the long-term signal variability (LTSV) VAD [26] are highly accurate for clean speech, but their performances in noisy environments are worse than that of the AFE VAD. A supervised VAD method based on a non-linear spectrum modification (NLSM) and GMM [27] is shown to be able to outperform both MFB and LTSV algorithms. In [3] a low-complexity VAD method based on *a posteriori* SNR weighted energy difference also shows a performance superior to several methods including the MFB VAD and the AFE VAD. These referred methods are all significantly superior to the G.729 [28] and G.723.1 [29] VAD algorithms included in their corresponding Voice-over-IP standards. The comparisons cited in the paragraph are all conducted on the Aurora 2 database [30].

This paper focuses on developing a VAD method that is unsupervised and able to generalize well to real-world data, irrespective of whether the data is corrupted by stationary or rapidly changing additive noise. In [3], *a posteriori* SNR weighted energy distance is used as the key speech feature for VAD and has demonstrated state-of-the-art performance in noisy environments, and the rational behind the feature is that the weighting of *a posteriori* SNR makes the weighted distance close to zero in non-speech regions for both clean and noisy speech. The VAD method in [3] first uses the *a posteriori* SNR weighted energy distance for selecting frames in a variable frame rate analysis and then conducts VAD based on the selected frames, and the method assumes the presence of speech within a certain period. While showing state-of-the-art performance, the method in [3] has several drawbacks. First, it makes an implicit assumption of speech presence, which does not always hold, and it is not uncommon that a VAD method assumes the presence of speech in every signal (e.g. a signal file in a database) [11], [14]. Secondly, the process of making VAD decisions through selecting frames first is cumbersome and suboptimal. Finally, estimating *a posteriori* SNR weighted energy distance is prone to noise. In the present work, we therefore propose a VAD method eliminating the assumption of speech presence by using pitch or spectral flatness as an anchor to find potential speech segments, directly using *a posteriori* SNR weighted energy distance without conducting frame selection, and applying speech enhancement. The proposed method differs from [3] in a number of ways: 1) two-stage denoising is proposed in order to enhance the noise robustness against both rapidly changing and relatively stationary noise, 2) pitch or spectral flatness is applied to detect high-energy noise segments and as an anchor to find potential speech segments, and 3) a segment based approach is used, in which VAD is conducted in segments, making it easier and more effective in determining a threshold for making decisions since a certain amount of speech and non-speech exists in each segment. The proposed method is called *robust voice activity detection* (rVAD).

Pitch information places an important role in rVAD and it has been used for existing VAD methods as well. For example, [31] combines pitch continuity with other speech features for detecting voice activity. In [32], long-term pitch divergence is used as the feature for VAD. The big difference here is that rVAD uses pitch as an anchor to locate potential speech regions, rather than as a VAD feature, and in rVAD the actual VAD boundaries are found by using *a posterior* SNR weighted energy distance as the feature. Concerning two-stage denoising, the concept has been applied in the literature although being different. In [33], two-stage denoising is conducted to handle reverberation and additive noise separately in a supervised fashion by training neural networks. A two stage Mel-warped Wiener filter approach is presented in [25]. In this work, the two-stage method aims to deal with high-energy noise in the first stage separately.

Most of the computation in rVAD lies in pitch detection, while the rest part of rVAD is computationally light. Therefore, we present a modified version of rVAD, where the time-consuming pitch detector is replaced by a computationally efficient spectral flatness (SFT) [34, 35, 36] detector. We call the modified algorithm *rVAD-fast*. We show that rVAD-fast is significantly (≈ 10 times) faster compared with rVAD with moderate degradation on VAD performance, which is beneficial when processing a larger amount of data or used on devices with computational constraints.

We demonstrate the performance of the rVAD method for voice activity detection on two databases consisting of various types of noise: the RATS [37] and Aurora-2 [30] databases. We further evaluate its performance for speaker verification on the RATS database as well as on the RedDots database [38] that we corrupt with additive noise. Experiment results show that rVAD is compared favourably with a number of existing methods.

The MATLAB source code of the proposed rVAD method (including rVAD-fast) is made publicly available [1], and the Python source code of rVAD-fast is also publicly available [2]. It is noted that a slightly different version of the rVAD source code - *rVAD1.0* - has already been made publicly available while no paper has been published to document the rVAD method and a number of studies have used it, covering applications such as voice activity detection in speaker verification [39, 40, 41], age and gender identification [42], emotion detection and recognition [43, 44, 45], and discovering linguistic structures [46]. A modified version is used for real-time human-robot interaction [47]. In addition, we have made the Python source code for training and testing of GMM-UBM and maximum *a posteriori* (MAP) adaptation based speaker verification publicly available [3].

The paper is organized as follows. The two-pass denoising method and the proposed rVAD method are presented in Sections 2 and 3, respectively. Section 4 describes rVAD-fast. Experimental results on the Aurora-2 database, the RATS database and the RedDots database are presented in Sections 5, 6 and 7, respectively. Finally, the paper is concluded in Section 8.

2. Robust VAD in noise

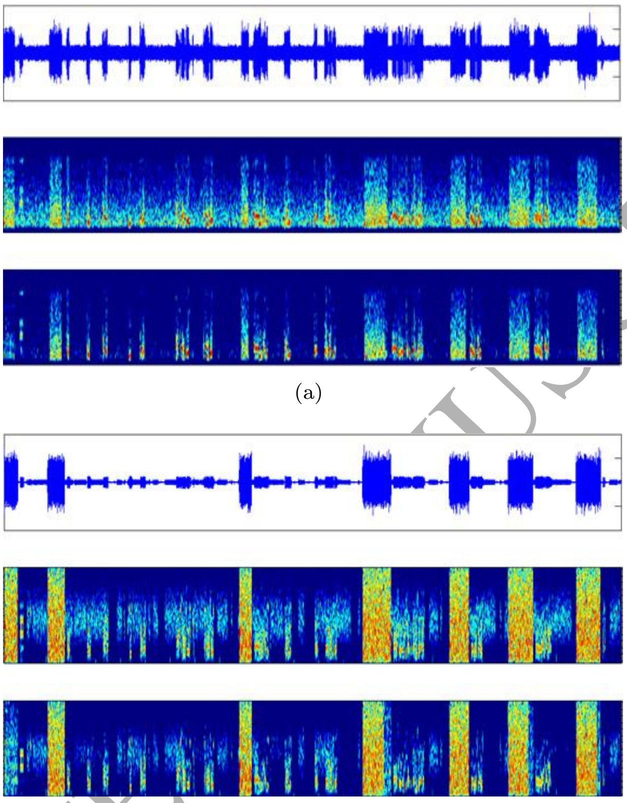
When robustness is of concern, a major challenge presented by real-world applications is that speech signals often contain both stationary noise as well as burst-like noise. It is very difficult to detect and remove burst-like noise since they have high energy and are rapidly changing and thus hard to estimate. To illustrate this, Fig. 1 shows, from the RATS database [8], two examples of noisy speech signals obtained by corrupting clean speech with different types of communication channel noise. The figure depicts waveform and spectrogram of the two noisy speech signals and spectrogram of their corresponding speech signals

[1] `http://kom.aau.dk/~zt/online/rVAD/rVAD2.0.zip; http://kom.aau.dk/~zt/online/rVAD/`

[2] `http://kom.aau.dk/~zt/online/rVAD/rVADfast_py_2.0.zip`

[3] `http://kom.aau.dk/~zt/online/rVAD/GMM-UBM_MAP_SV_Python.zip`

denoised by the minimum statistics noise estimation (MSNE) [48] based spectral subtraction. It is noticed that high-energy noise is largely kept intact after denoising, except for at the beginning of each utterance where the noise estimate is close to the real noise since there is only high-energy noise in the beginning of an utterance. We further did preliminary VAD experiments using a classical statistical method [19] on a few files and the performance is not encouraging (with over 20% frame error rate). These initial tests show that burst-like noise presents a significant challenge to both denoising and VAD methods.


(a)

(b)

Figure 1: *Noisy and denoised speech: (a) Noisy speech from Channel A: waveform of noisy speech (the first panel), spectrogram of noisy speech (the second panel) and spectrogram of denoised speech (the third panel); (b) noisy speech from Channel H with the same order of panels as in (a).*

Due to the very different characteristics of stationary and burst-like noise, the two types of noise require different types of processing. To deal with this problem, we devise a two-pass denoising method, as detailed in the presentation of rVAD in Section 3. In the first pass, the *a posteriori* SNR weighted energy difference measure [3] is used to detect high-energy segments. If a high-energy segment is detected and it does not contain pitch, the segment is considered non-speech and its samples are set to zero.

After detecting and denoising high-energy noise segments, in the second pass, a general speech enhancement method is applied to the first-pass-denoised speech signal in order to remove the remaining noise that is relatively more stationary. In the classical speech enhancement framework [49], accurate noise estimation is important and the widely used methods include minimum statistics noise estimation (MSNE) [48], minimum mean-square error (MMSE) [50, 51], and minima controlled recursive averaging (MCRA) [52]. Based on an estimate of additive-noise spectrum, a spectral subtraction (SS) method [53] is then used to subtract the estimate from the noisy speech spectrum. Another interesting speech enhancement method is the one in the advanced front-end [25], which aims at improving speech recognition in noisy environments. In [54] it is found that the AFE enhancement method outperforms MMSE-based methods for noise-robust speech recognition. Recently DNN based speech enhancement methods have also been proposed for improving speech intelligibility [55], automatic speech recognition [56] and speaker verification [57, 58].

Unsupervised VAD methods often rely on a predefined or adaptive threshold for making VAD decisions, and finding this threshold is a challenging problem in noisy environment. Therefore, it is not uncommon that a VAD method assumes the presence of speech in every signal (e.g. a signal file in a database) that the method is applied upon [11], [14]. This assumption holds for benchmark speech databases, but does not in real-world scenarios where it is possible to have no speech in a long duration. As we know, a speech signal must contain pitch, which motivates us to propose to use pitch (or its replacement speech flatness) as an anchor or indicator for speech presence, namely speech is present in a signal if pitch is detected. This leads to the VAD process in this work: detecting extended pitch segments first and then detecting speech activity within the segments. Extended pitch segment detection plays a key role in rVAD. First, this provides an anchor to locate speech segments. Secondly, it enables to exclude a substantial amount of non-speech, potentially noisy, part from a speech signal. Finally, this results in a segment-based approach in which voice activity detection operates in segments, making it easier and more effective in terms of determining a decision threshold as both speech and non-speech are guaranteed to be present in each segment.

3. rVAD: an unsupervised segment-based VAD method

The block diagram of the proposed rVAD method is shown in Fig. 2. It consists of the following steps: the first pass denoising (high-energy segment detection, pitch based noise-segment classification, and setting high-energy noise segments to zero), the second pass denoising, extended pitch segment detection, and the actual VAD. These steps are detailed in this section.

The noise-corrupted speech signal is modelled using the additive noise signal model as

$$x(n) = s(n) + v(n) \tag{1}$$

where $x(n)$ and $s(n)$ represent the noisy and clean speech sample at time n, respectively, and $v(n)$ the sample of additive noise at time n. The signal $x(n)$ is first filtered by a high-pass filter to remove the DC component and low frequency noise. A first-order high-pass filter with a cutoff frequency of 60 Hz is applied to remove low-frequency noise. For simplicity, this is not included in the equations of this paper.

To conduct short-time speech analysis, the signal is partitioned into frames of 25 ms in length with a frame shift of 10 ms, without using pre-emphasis or a hamming window unless stated otherwise.

3.1. The first pass denoising

In the first pass, high-energy segments are detected by using an *a posteriori* SNR weighted energy difference measure [3] as follows:

(a) Calculate the *a posteriori* SNR weighted energy difference of two consecutive frames as

$$d(m) = \sqrt{|e(m) - e(m-1)| * \max(SNR_{post}(m), 0)} \tag{2}$$

where m is the frame index, $e(m)$ the energy of the m^{th} frame of noisy speech $x(n)$, and $SNR_{post}(m)$ is *a posteriori* SNR that is calculated as the logarithmic ratio of $e(m)$ to the estimated energy of the m^{th} frame of noise signal $v(n)$:

$$SNR_{post}(m) = 10 * log_{10} \frac{e(m)}{\tilde{e}_v(m)} \tag{3}$$

In Eq.(2), the square root is taken to reduce the dynamic range, which differs from [3] where the square root is not applied. Energy is calculated as the sum of the squares of all samples in a frame. The noise energy $\tilde{e}_v(m)$ is estimated as follows. First, the noisy speech signal $x(n)$ is partitioned into super-segments of 200 frames each (about 2s): $x(p) = s(p) + v(p)$, $p = 1, \dots, P$, where P is the number of super-segments in an utterance. For each super-segment $x(p)$, the noise energy $e_v(p)$ is calculated as the energy of the frame ranked at 10% of lowest energy within the super-segment. Thereafter, the noise energy $\tilde{e}_v(p)$ is calculated as the smoothed version of $e_v(p)$ with a forgetting factor of 0.9 as follows:

$$\tilde{e}_v(p) = 0.9 * \tilde{e}_v(p-1) + 0.1 * e_v(p). \tag{4}$$

Noise energy of the m^{th} frame, $\tilde{e}_v(m)$, takes the energy value $\tilde{e}_v(p)$ of the p^{th} super-segment which the m^{th} frame belongs to.

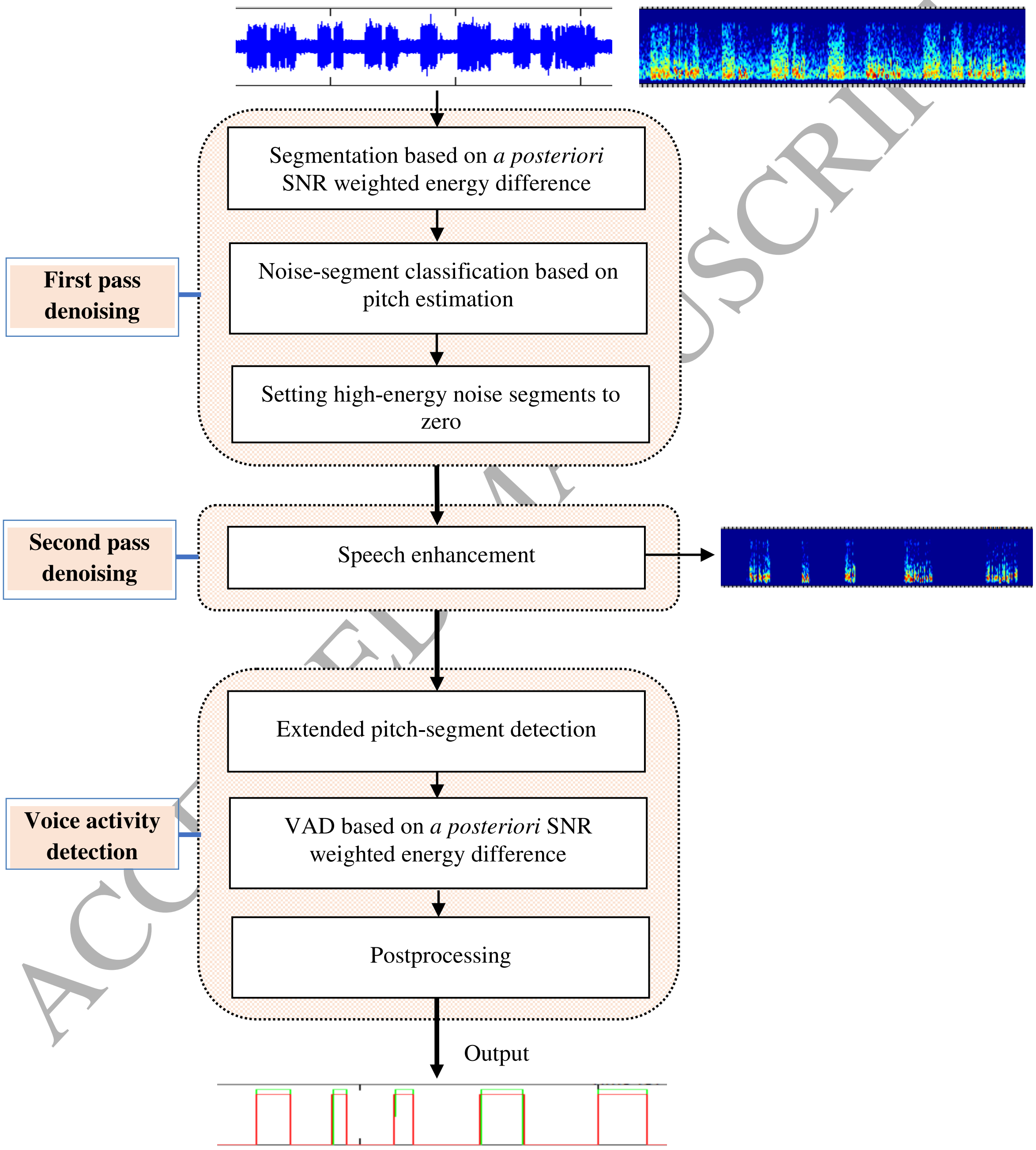

Segmentation based on *a posteriori* SNR weighted energy difference
Noise-segment classification based on pitch estimation
Setting high-energy noise segments to zero
First pass denoising
Second pass denoising
Speech enhancement
Extended pitch-segment detection
VAD based on *a posteriori* SNR weighted energy difference
Postprocessing
Voice activity detection
Output

Figure 2: *Block diagram of the rVAD method.*

(b) Central-smooth the *a posteriori* SNR weighted energy difference,

$$\bar{d}(m) = \frac{1}{2N+1} \sum_{i=-N}^{N} d(m+i) \tag{5}$$

where $N = 18$.

(c) Classify a frame as a high-energy frame if $\bar{d}(m)$ is greater than a threshold $\theta_{he}(m)$. For each super-segment p (containing 200 frames), $\theta_{he}(p)$ is computed as follows:

$$\theta_{he}(p) = \alpha * max\big\{e((p-1)*200+1), \ldots, e(m) \ldots, e(p*200)\big\} \tag{6}$$

where $\alpha = 0.25$. $\theta_{he}(m)$ takes the threshold value $\theta_{he}(p)$ of the p^{th} super-segment which the m^{th} frame belongs to. Alternatively, the threshold can be calculated recursively using a forgetting factor.

(d) Consecutive high-energy frames are grouped together to form high-energy segments.

(e) Within a high-energy segment, if no more than two pitch frames are found, the segment is classified as noise, and the samples of the segment are set to zero.

Motivation of this pass is two-fold: first, to avoid overestimating noise due to the burst-like noise when applying a noise estimator in the second pass denoising and secondly, to detect and denoise high-energy non-speech parts, which are otherwise difficult for conventional denoising and VAD methods to deal with.

3.2. The second pass denoising

Any speech enhancement method is applicable for the second pass denoising. Three spectral subtraction based speech enhancement methods are considered in this work, and they rely on different noise estimation approaches: MMSE [51], MSNE [48], and a modified version of MSNE (MSNE-mod).

The unbiased noise power estimation in the conventional MSNE can be expressed as,

$$\hat{\lambda}_v(m,k) = B_{min} * min\big\{P(m,k), P(m-1,k), \ldots, P(m-l,k)\big\} \tag{7}$$

where m is the frame index, k the frequency bin index, $B_{min}(m,k)$ the bias compensation factor, $P(m,k)$ the recursively smoothed periodogram, and l the length of the finite window for searching the minimum.

In the proposed MSNE-mod, noise estimate $\hat{\lambda}_v(m,k)$ is not updated during the detected high-energy noise segments (which are set to zero). Besides, if more than half of the energy is located within the first 7 frequency bins ($< 217\ Hz$), the values of the 7 frequency bins are set to zero to further remove low-frequency noise in addition to the use of the first-order high-pass filter mentioned earlier.

3.3. Extended pitch segment detection

The VAD algorithm is based on the denoised speech and the pitch information generated from the previous steps. The fundamental assumption is that all speech segments should contain a number of speech frames with pitch. In this algorithm, pitch frames are first grouped into pitch segments, which are then extended from both ends by 60 frames (600ms), based on speech statistics, in order to include voiced sounds, unvoiced sounds and potentially non-speech parts. This strategy is taken for the following reasons: 1) pitch information is already extracted in the previous steps, 2) pitch is used as an anchor to trigger the VAD process and 3) many frames can be potentially discarded if an utterance contains a large portion of non-speech segments, which are also non-pitch segments.

3.4. Voice activity detection

The *a posteriori* SNR weighted energy difference is applied now to each extended pitch segment to make VAD decision as follows:

(a) Calculate *a posteriori* SNR weighted energy difference $d^{'}(m)$ according to Eq. (2). To calculate $SNR_{post}(m)$ using Eq. (3), the noise energy $\tilde{e}_v^{'}(m)$ is estimated as the energy ranked at 10% of lowest frame energy within the extended pitch segment.

(b) Central-smooth the *a posteriori* SNR weighted energy difference with $N = 18$ as in Eq.(5), resulting in $\bar{d}^{'}(m)$.

(c) Classify a frame as speech if $\bar{d}^{'}(m)$ is greater than the threshold (θ_{vad})

$$\theta_{vad} = \beta * \frac{1}{L} \sum_{j=1}^{L} \bar{d}^{'}(j) \tag{8}$$

where L is the total number of frames with pitch in the extended pitch segment and the default value for β is set to 0.4.

(d) Apply post-processing. The assumptions here are speech frames should not be too far away from its closest pitch frame, and within a speech segment, there should be a certain number of speech frames without pitch. While it is possible to gain improvement by analysing noisy speech as well, we analyse only a few clean speech files that are not included in any experiments in this paper and then derive and apply the following rules. First, frames that are 33 frames away from the pitch segment to the left and 47 frames away to the right are classified as non-speech, regardless the VAD results above, which covers 95% of the cases based on the a few speech files. On the other hand, frames that are within 5 frames to the left and 12 frames to the right of the pitch segments are classified as speech, again regardless the VAD results, which leaves out 5% of the cases based on the a few speech files. The concept is sort of similar to that of hangover schemes used in some VAD methods. Furthermore, segments with energy below 0.05 times the overall energy is removed.

4. rVAD-fast based on spectral flatness

As the computational complexity of pitch detection is relatively high, we investigate alternative measures to pitch, for example, spectral flatness (SFT) [34, 35, 59]. Our primitive study shows that SFT is a good indicator to tell whether or not there is pitch in a speech frame. Replacing the pitch detector by a simple SFT based voiced/unvoiced speech detector, leads to a more computationally efficient algorithm called rVAD-fast. To extract the SFT feature, a hamming window is first applied to a speech frame before taking short-time Fourier transform (STFT). After STFT, the signal is represented in the spectral domain as

$$X(m, k) = S(m, k) + V(m, k). \tag{9}$$

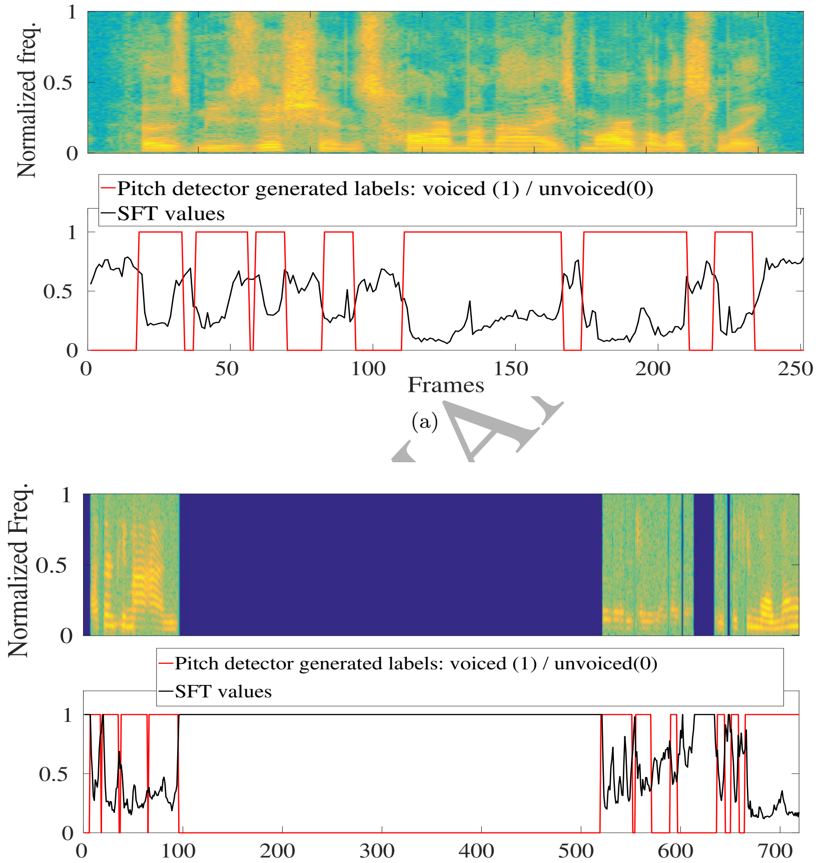
Thereafter, SFT is calculated as

$$SFT(m) = \frac{exp(\frac{1}{K}\sum_{k=0}^{K-1} ln\ |X(m,k)|)}{\frac{1}{K}\sum_{k=0}^{K-1} |X(m,k)|} \tag{10}$$

where $|X(m, k)|$ denotes the magnitude spectrum of k^{th} frequency bin for the m^{th} frame, and K is the total number of frequency bins. As SFT is used as a replacement for pitch in this work, SFT values are compared against a predefined threshold θ_{sft} to decide whether their corresponding frame is voiced or unvoiced.

Figures 3(a) and 3(b) illustrate spectrogram, pitch labels (1 for pitch and 0 for no pitch) and SFT values of a speech signal from TIMIT (clean) and those of a signal from the NIST 2016 SRE evaluation (noisy), respectively. The figures show that if we choose a θ_{sft} value of 0.5 (i.e. if SFT $\leq$ 0.5, a frame is said to contain pitch), the labels generated by SFT are close to those generated by the pitch detector.

We extensively studied the effect of different threshold values θ_{sft} on the performance of SFT as a replacement of the pitch detector, which will not be detailed in this paper. Briefly, we compared the output of the SFT detector of different values of θ_{sft} with the output of the pitch detector on large number of utterances from various databases including NIST 2016 SRE (evaluation set) [60], TIMIT [61], RSR2015 [62], RedDots [38], ASVspoof2015 [63] and noisy versions (car, street, market, and white) of ASVspoof2015. It is observed that a threshold value of 0.5 gives the best match between SFT and pitch, and this value is used for experiments in this paper.



(a)

(b)

Figure 3: *Spectrogram, pitch labels (1 for pitch and 0 for no pitch) and SFT values of a speech signal from (a) TIMIT (b) the NIST 2016 SRE evaluation set.*

5. Experiments on the Aurora-2 database

To evaluate the performance of the proposed rVAD method, experiments are conducted on a number of databases and for different tasks. In this section, we compare rVAD with ten existing VAD methods (both supervised and unsupervised) in terms of VAD performance on the test sets of the Aurora-2 database. Aurora-2 [30] has three test sets A, B and C, all of which contain both clean and noisy speech signals. The noisy signals in Set A are generated from clean data by using a filter with a G.712 characteristic and mixing four noise types including subway, babble, car and exhibition with SNR values ranging across 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, and -5 dB. Set B is created similarly with the only difference being that the types of noise are restaurant, street, airport and train station. In Set C, clean speech is corrupted by subway and street noise, in addition to a Motorola integrated radio systems (MIRS) characteristic filter being applied instead of that of G.712.

The reference VAD labels for the Aurora-2 database are generated with the HTK recognizer [64] which is trained using the training set of Aurora-2. Whole word models are created for all digits. Each of the whole word digit models has 16 HMM states with three Gaussian mixtures per state. The silence model has three HMM states with six Gaussian mixtures per state. A one state short pause model is tied to the second state of the silence model. The speech feature consists of 12 MFCC coefficients, logarithmic energy as well as their corresponding Δ and $\Delta\Delta$ components. In [65], it is confirmed that forced-alignment speech recognition is able to provide accurate and consistent VAD labels, matching closely transcriptions made by an expert labeler and being better than most of those made by non-expert labelers. The generated reference VAD labels are made publicly available [4].

Several metrics are used to characterize VAD performance. The frame error rate (FER) is defined as

$$FER = 100 * \frac{\# \left\{false\ rejection\ frames\ +\ false\ alarm\ frames\right\}}{\#\ total\ frames} \tag{11}$$

False alarm rate P_{fa} is the percentage of non-speech frames being misclassified as speech and miss rate P_{miss} is the percentage of speech frames being misclassified as non-speech. Detection cost function (DCF) is defined as

$$DCF = (1-\gamma) * P_{miss} + \gamma * P_{fa} \tag{12}$$

where the weight γ is equal to 0.25, which penalizes missed speech frames more heavily.

Through out this paper, the default configuration for rVAD is the one that includes two-pass denoising (with the second being MSNE) and the post-processing, which is also the default configuration in the released rVAD source code, and this configuration, shown in italic font in tables, is used for all experiments in this paper unless stated otherwise. We do not change the settings and parameters while testing rVAD on different datasets so as to evaluate its generalization ability, unless a test is specifically for assessing the effects of changing settings, e.g. the threshold β in Eq. (8).

For experiments conducted in this paper, pitch extraction is realized using the pitch estimator in [66], unless stated otherwise. It has been experimentally shown that using Praat pitch extraction [67] gives almost the same VAD results.

5.1. Comparison with referenced methods and evaluation of different configurations

Table 1 presents the VAD results averaged over the three test sets of Aurora-2, which accounts for 70070 speech files, for various methods. VQVAD [11] first applies a speech enhancement method and an energy VAD to a testing utterance in order to automatically label a small subset of MFCCs as speech or non-speech; afterwards, these MFCCs are used to train speech and non-speech codebooks using k-means and all the frames in the utterance are labeled using nearest-neighbor classification. Sohn *et al.* VAD [19] is an unsupervised VAD method based on a statistical likelihood ratio test, which uses Gaussian models to represent the distributions of speech and non-speech energies in individual frequency bands. The method also uses a hangover scheme. The VoiceBox toolkit [5] implemented version of Sohn *et al.* VAD is used in this study. *Kaldi* VAD is the the Kaldi toolkit's [68] energy based VAD for which we use the default parameters (-vad-energy-threshold=5.5, –vad-energy-mean-scale=0.5) as included in the SRE16 script. Note that Kaldi is a widely used open source software for speaker and speech recognition.

Results for the G.729, G.723.1 and MFB VAD methods are cited from [17], results for the LTSV and GMM-NLSM methods are from [27], and results for the DSR-AFE and variable frame rate (VFR) methods are from [3]. The comparison in this table is conducted in terms of frame error rate (FER) since results of

[4] `http://kom.aau.dk/~zt/online/rVAD/`

[5] `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`

LTSV and GMM-NSLM are only available in terms of FER. Note that the identical experimental settings and labels are used across [3, 17, 27] and the present work, so the comparison is valid.

The results in Table 1 clearly show that rVAD with the default configuration gives significantly lower average FER than those of the compared VAD methods. It outperforms all referenced VAD methods, with big margins, under all SNR levels including the Clean condition. The closest one is the VFR VAD [3], which is our previous work that also uses *a posteriori* SNR weighted energy distance as the feature for VAD decision. The GMM-NLSM [27] VAD provides good performance as well, but still with a 3% (absolute) higher FER as compared with rVAD, and furthermore it should be noted that GMM-NLSM is a *supervised* VAD where the GMMs are trained using multicondition training data of the Aurora 2 database. The next one in line is the VAD method in the DSR AFE frontend [25], which is an unsupervised VAD and gives a more than 5% (absolute) higher FER than that of rVAD. DSR AFE performs well under low SNRs, for example at 0dB and -5dB, with FERs close to those of rVAD; from 5dB and above, however, its performance is far below that of rVAD. The remaining compared methods have even much higher FERs. For example, VQVAD[11] gives a 36% FER, as compared with a 11% FER achieved by rVAD.

Table 1: *Comparison of rVAD (with the default configuration) with other methods on the test datasets of Aurora-2.*

Methods	FER (%)							Avg.
	Clean	20 dB	15 dB	10 dB	5dB	0 dB	-5 dB	FER
VQVAD[11]	17.26	33.64	36.34	38.77	41.05	43.03	46.07	36.59
G.729 [28]	12.84	24.53	26.13	27.38	29.13	32.23	35.21	26.78
Sohn et al. [19]	17.37	20.16	21.97	24.48	27.96	33.12	39.76	26.40
Kaldi Energy VAD [68]	9.88	26.54	26.61	26.62	26.62	26.62	26.62	24.22
G.723.1 [29]	19.45	21.31	23.29	24.44	26.30	26.56	28.58	23.99
MFB [17]	6.92	15.39	17.70	20.12	22.75	26.16	31.09	20.02
LTSV [26]	9.50	15.90	16.80	18.20	21.00	26.10	28.80	19.50
DSR AFE [25]	18.41	15.16	14.96	14.59	14.54	15.62	22.08	16.48
GMM-NLSM (supervised) [27]	10.95	11.20	11.43	11.73	13.35	17.44	23.52	14.23
VFR [3]	8.10	8.30	9.00	10.60	13.50	19.50	28.20	13.90
rVAD (default)	**6.90**	**7.30**	**7.64**	**8.43**	**11.09**	**16.01**	**21.48**	**11.26**

Table 2 compares the results of various configurations for rVAD. It is shown that when evaluating on the Aurora-2 database, the first-pass denoising of rVAD does not make difference, for which the reason is that burst-like high-energy noise prominently present in the RATS database has much less presence in Aurora-2 while the aim of the first pass denoising is to remove burst-like noise. The second pass denoising with MSNE or MMSE is able to boost the performance of rVAD and there is almost no performance difference between the two enhancement methods. MSNE-mod, however, does not perform as well as MSNE, for which the reason is that MSNE-mod is tailored to the special noise characteristics of the RATS database. The postprocessing step is also shown to be important for improving the performance of rVAD.

Table 2: *Comparison of rVAD (with different configurations) with other methods on the test datasets of Aurora-2.*

Methods	Denoising		FER (%)							Avg.
	1st-pass	2nd-pass	Clean	20 dB	15 dB	10 dB	5dB	0 dB	-5 dB	FER
rVAD (default)	✓	MSNE	6.90	7.30	7.64	**8.43**	11.09	16.01	21.48	11.26
(two-pass denoising)	✓	MMSE	6.96	7.35	7.75	8.58	11.02	15.93	21.55	11.30
	✓	MSNE-mod	6.86	7.07	**7.44**	8.68	13.13	19.73	25.56	12.63
rVAD (w/o denoising)	×	×	6.89	7.37	7.69	8.81	13.07	19.05	23.23	12.30
rVAD (w/o 1st-pass	×	MSNE	6.89	7.30	7.64	8.45	11.01	15.88	21.44	**11.23**
denosing)	×	MMSE	6.95	7.35	7.76	8.60	11.00	15.85	21.51	11.28
	×	MSNE-mod	**6.85**	**7.06**	**7.44**	8.66	13.07	19.59	25.52	12.59
rVAD (w/o postproc)	✓	MSNE	8.82	9.53	9.78	10.05	11.32	15.47	21.59	12.37

Overall, the experimental results demonstrate the effectiveness of the rVAD method on the Aurora-2 database, which is a very different database from the RATS database that rVAD was originally devised for. This confirms the good generalization ability of rVAD.

5.2. Sensitivity of rVAD with changing threshold β on VAD performance

In rVAD, β in Eq. (8) is an important thresholding parameter that controls the aggressive level of rVAD; the larger the value of β, the higher the aggressive level. Table 3 presents VAD results, in terms of false rejections rate P_{miss}, false alarms rate P_{fa} and FER, of rVAD with various β values. Apart from varying β values, the default configuration of rVAD is used. For small β values, P_{miss} is small while P_{fa} is large and vice versa. As of FER, a threshold value of 0.4 gives the best performance while a change of ± 0.1 in β only marginally change the performance. The results confirm the stability of rVAD with respect to changing this threshold.

In Table 3, we additionally compare rVAD with several other methods now also in terms of P_{miss} and P_{fa} while in Table 1 only FER performance is compared. This allows us to compare rVAD with other methods in terms of P_{fa}, while fixing P_{miss} or the other way around. Table 3 clearly shows that rVAD outperforms the referenced methods with big margins.

Table 3: *Performance of rVAD with various threshold β values in comparison with several referenced methods on the Aurora-2 database.*

Method	Threshold (β)	P_{miss}(%)	P_{fa}(%)	Avg. FER (%)
rVAD	0.1	1.21	52.90	14.97
	0.2	2.16	42.43	12.87
	0.3	3.21	35.11	11.70
	0.4 (default)	4.86	28.91	11.26
	0.5	7.54	23.59	11.81
	0.6	11.41	19.03	13.44
	0.7	16.48	15.25	16.15
VQVAD [11]	-	7.02	47.32	36.59
Sohn et al. [19]	-	17.31	51.48	26.40
Kaldi Energy VAD [68]	-	1.53	86.75	24.22
DSR AFE [25]	-	3.32	56.61	16.48

5.3. Evaluation of rVAD-fast

Table 4 compares the performance of the rVAD-fast method with rVAD on the test datasets of Aurora-2 in terms of FER and processing time. It can be seen that rVAD-fast is approximately an order of magnitude faster than rVAD at the cost of moderate performance degradation. This presents an advantage when processing a large number of speech files or running on low-resource devices. For measuring the processing time, the algorithms were run on a desktop computer with Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz and 16 GB RAM, and we measured the CPU time.

Table 4: *Comparison of rVAD-fast with rVAD for voice activity detection on the test datasets of Aurora-2 in terms of FER and average CPU processing time*

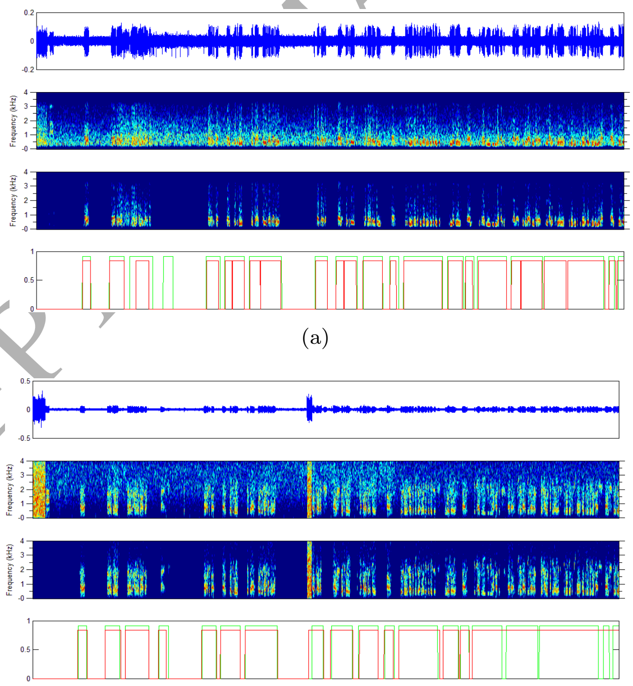
Methods	FER (%)								Avg. time	Times as fast
	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Avg.	sec/file	rVAD-fast
rVAD (MSNE, default)	6.90	7.30	7.64	8.43	11.09	16.01	21.48	11.26	0.1253	
rVAD-fast(MSNE)	7.25	8.75	9.15	10.07	12.37	17.35	25.18	12.87	0.0085	≈ 14×
rVAD(MMSE)	6.96	7.35	7.75	8.58	11.02	15.93	21.55	11.30	0.1181	
rVAD-fast(MMSE)	7.26	8.78	9.22	10.20	12.61	17.71	25.95	13.10	0.0068	≈ 17×
rVAD(MSNE-mod)	6.86	7.07	7.44	8.68	13.13	19.73	25.56	12.63	0.1206	
rVAD-fast(MSNE-mod)	7.25	8.71	9.30	11.29	15.31	21.01	30.40	14.75	0.0135	≈ 9×

6. Experiments on the RATS database

In this section, rVAD is evaluated against nine different VAD methods (both supervised and unsupervised) on the RATS database [37, 69] and it consists of audio recordings selected from several existing and new data sources. All recordings are re-transmitted through 8 different noisy communication channels, labelled by the letters A through H. The database is for evaluating methods such as voice activity detection, speaker identification, language identification and keyword spotting.

6.1. Illustrative VAD and denoising results of rVAD

Figure 4 illustrates results of rVAD for two speech recordings taken from the RATS database. It shows that rVAD is able to remove both stationary and burst-like noise and performs well in terms of VAD.



(a)

(b)

Figure 4: *Illustrative results of the proposed rVAD method on speech signals from the RATS database: (a) noisy speech from Channel A with four panels presenting waveform of the original signal, its spectrogram, spectrogram of the denoised signal and VAD results where green and red colours represent true labels and algorithm outputs, respectively; b) noisy speech from Channel H with the same order of panels as in (a).*

6.2. VAD results

Data used for this evaluation contains 8x25 = 200 speech files randomly selected from the RATS database and it amounts to approximately 44 hours of speech in total, which is a sufficient quantity for evaluating VAD methods. Table 5 compares rVAD of default configuration with other VAD methods.

Table 5 shows that rVAD gives substantially lower FER compared with Sohn *et al.*, Kaldi [68], VQVAD [11] and SSGMM [12] VAD methods. SSGMM [12] is neighter included in the experiments for the Aurora-2 database nor for the RedDots database, as SSGMM does not work for short utterances (due to the data sparsity issue caused by short utterances when training GMMs), as also stated in [12], and it produces empty label files for Aurora-2 or RedDots utterances. Kaldi VAD provides the lowest value of P_{miss}, but with significantly higher P_{fa} and FER values. $P_{fa} = 86\%$ indicates that Kaldi VAD classifies most frames as speech. Apparently, simple energy-threshold based VAD methods do not work well under highly noisy environments. VQVAD, a widely used and well-performing VAD method in the speaker verification domain, provides as high as 36% FER. It is interesting to note that VQVAD and Sohn *et al.* achieve 36% and 26% FER, respectively, on both the Aurora-2 database and the RATS database, as shown in Tables 1 and 5.

Table 5: *Comparison of VAD performance of the proposed rVAD of default configuration with other methods on the RATS database.*

Methods	P_{miss}(%)	$P_{fa(\%)}$	FER (%)
Kaldi Energy VAD [68]	3.10	85.98	51.62
DSR AFE [25]	4.17	76.24	46.36
VQVAD[11]	41.65	32.85	36.49
Sohn et al. [19]	30.36	24.51	26.93
SSGMM[12]	30.36	13.24	20.33
rVAD (default)	17.00	5.52	**10.27**

Concerning the different configurations of rVAD, Table 6 shows the first-pass denoising improves the performance slightly. The modified MSNE, which is tailored towards the RATS database, boosts the VAD performance modestly while MSNE and MMSE does not. The best performance is given by first-pass denoising and modified MSNE based second-pass denoising. Overall, the improvement brought by denoising is modest, for which the reason could be that the voice activity detection is conducted on the basis of the extended pitch segments in which the non-speech segments that do not contain pitch have been excluded during the process of detecting pitch segments.

Table 6: *Comparison of VAD performance of the proposed rVAD of different configurations on the RATS database.*

Methods	Denoising		P_{miss}(%)	P_{fa}(%)	FER (%)
	1st-pass	2nd-pass			
rVAD (default)	✓	MSNE	17.00	5.52	10.27
(with two-pass denoising)	✓	MMSE	16.67	5.84	10.32
	✓	MSNE-mod	17.63	4.82	**10.12**
rVAD (w/o denoising)	×	×	16.30	6.59	10.61
rVAD (w/o 1st-pass	×	MSNE	17.88	5.63	10.71
denoising)	×	MMSE	18.21	5.56	10.80
	×	MSNE-mod	17.52	5.16	10.28
rVAD (w/o postproc)	✓	MSNE	18.07	8.29	12.34

We study the sensitivity of threshold β in Eq. (8) of rVAD for voice activity detection. Table 7 presents the results on the RTAS database when using different values of β; other than this, the default configuration of rVAD is used. It is observed that increasing the value of the VAD threshold β makes rVAD more aggressive, i.e. increased P_{miss} and decreased P_{fa}, and thus less frames are classified as speech. The β value should be chosen according to the application in hand. Furthermore, the results show that rVAD performs well in a wide range of values of threshold β, demonstrating its advantage of stability.

Table 7: *Performance of rVAD with various threshold β values on the RATS database.*

Method	Threshold (β)	P_{miss}(%)	P_{fa}(%)	FER (%)
rVAD (MSNE)	0.1	8.85	10.58	9.86
	0.2	11.36	8.16	9.48
	0.3	14.04	6.60	9.68
	0.4(default)	17.00	5.52	10.27
	0.5	20.33	4.70	11.17
	0.6	24.16	4.02	12.36
	0.7	28.79	3.44	13.94

6.3. VAD results for NIST 2015 OpenSAD challenge

In [15], the rVAD method is compared with several methods on the NIST 2015 OpenSAD challenge [70] that is also based on the RATS database. Table 8 cites results for several methods from [15], our joint work with several other teams. The GMM-MFCC and GMM-PNCC methods are similar to [11], but using GMMs trained with maximum-likelihood instead of codebooks, which leads to some improvement. PNCCs [71] are known to be robust against noise. The i-vector based VAD is a supervised VAD [21]. rVAD shows a performance substantially better than those of Sohn *et al.*, GMM-MFCC and GMM-PNCC, and a performance very close to that of the supervised i-vector method trained on the NIST 2015 OpenSAD challenge.

Table 8: *VAD results [15] of several methods on Dev data of NIST 2015 OpenSAD challenge.*

Methods	P_{fa}(%)	P_{miss}(%)	DCF
GMM-MFCC [15]	6.15	43.17	0.1540
GMM-PNCC [15]	7.72	17.14	0.1008
Sohn et al. [19]	6.35	38.89	0.1449
i-vector (supervised) [21]	2.77	10.09	**0.0460**
rVAD (MSNE-mod)	4.78	5.75	0.0502

6.4. Speaker verification results on the RATS database

In [72], the rVAD method was applied as a preprocessing step of text-independent speaker verification (TI-SV) systems built for the RATS database under the DARPA RATS project. The TI-SV systems use 60-dimension MFCCs as features and 600-dimension i-vector [73] together with probability linear discriminate analysis (PLDA) based scoring as the speaker verification back-end. Praat pitch extration [67] was used for the rVAD method in this experiment. rVAD was first compared with a neural network based supervised VAD method developed by Brno University of Technology (BUT-VAD) [74]. The neural network deployed in BUT-VAD has 9 outputs for speech and 9 outputs non-speech, each of which corresponds to one of the 9 channels (one source and 8 retransmitted). The outputs are smoothed and merged into speech and non-speech segments using a hidden Markov model (HMM) with Viterbi decoding. The neural network was trained on RATS data defined for the VAD task. Another supervised VAD method is GMM-PLP-RASTA where perceptual linear predictive (PLP) coefficients with RASTA-based [75] cepstral mean normalization being applied are used as the feature and two 2048-component GMMs (one for speech and another for non-speech), trained on RATS, are used as the models for VAD.

Evaluation was conducted on a development set of 30s-30s enrolment and test condition and the evaluation criterion for speaker verification is equal error rate (EER) [72]. The unsupervised rVAD (MSNE-mod) method, the supervised BUT-VAD and GMM-PLP-RASTA methods[74] yield EERs of 5.6%, 5.4%, 6.7%, respectively. rVAD (MSNE-mod) performs marginally worse than the supervised BUT-VAD trained on the RATS database, but better than the supervised GMM-PLP-RASTA also trained on the RATS database. For more details about the systems see [72].

7. Experiments on the RedDots database for TD-SV

In this section, we compare rVAD and rVAD-fast with other VAD methods in the context of text-dependent speaker verification (TD-SV) on the male part-01 task of the RedDots 2016 challenge database [38], which consists of short utterances and is one of most used databases for speaker verification. The database was collected in many countries and mostly in office environments, hence introducing great diversity in terms of speakers and channels. There are 320 target models for training and each target has 3 speech signals/sessions for building their particular model. For the assessment of TD-SV, there are three types of non-target trials: target-wrong (TW) $(29, 178)$, impostor-correct (IC) $(120, 086)$ and impostor-wrong (IW) $(1, 080, 774)$.

The modelling method for SV adopts Gaussian mixture model-universal background model (GMM-UBM) since GMM based methods are known to outperform the i-vector technique for SV with short utterances. MFCCs of 57 dimensions (including static, Δ and $\Delta\Delta$) with RASTA filtering [75] are extracted from speech signals with a 25 ms hamming window and a 10 ms frame shift. After VAD, detected speech frames are normalized to have zero mean and unit variance at utterance level. A gender independent GMM-UBM, consisting of 512 mixtures and having diagonal covariance matrices, is trained using non-target data from TIMIT over 630 speakers (6300 utterances). Target models are derived from the GMM-UBM with 3 iterations of MAP adaptation (with relevance factor 10.0 and only applied to Gaussian mean vectors) using the training data of the particular target model. During the test, feature vectors of a test utterance is scored against the target model and GMM-UBM to calculate log likelihood ratio. Table 9 presents the TD-SV performance for different methods which are generally proposed for the SV.

Table 9: *Performance of VAD methods for text-dependent speaker verification on RedDots part-01 (male)*

Methods	[%EER/minDCFx100]			Average
	Target-wrong	Impostors-correct	Impostors-wrong	(%EER/minDCF)
	(TW)	(IC)	(IW)	
no VAD	6.36/3.071	2.82/1.400	1.26/0.525	3.48/1.665
Kaldi Energy VAD [68]	6.26/2.662	3.62/1.625	1.67/0.618	3.85/1.635
Sohn et al. [19]	4.78/2.492	**2.40/1.200**	1.01/0.295	2.73/1.329
VQVAD [11]	**3.70**/1.652	2.94/1.520	0.89/0.284	**2.51**/1.152
rVAD(MSNE, default)	3.79/1.572	2.93/1.328	1.01/0.273	2.58/1.058
rVAD(MMSE)	4.16/1.523	3.02/1.329	**0.92**/0.300	2.70/1.050
rVAD(MSNE-mod)	3.82/**1.498**	2.83/1.290	0.98/0.284	2.54/**1.024**
rVAD-fast(MSNE)	4.10/1.682	2.86/1.228	**0.92/0.272**	2.63/1.061
rVAD-fast(MMSE)	5.09/2.242	2.74/1.276	1.14/0.383	2.99/1.300
rVAD-fast(MSNE-mod)	3.97/1.705	2.64/1.225	0.95/0.283	2.52/1.071

The experiment results in Table 9 show that that all VAD methods (except for Kaldi VAD) are able to outperform the system without VAD either in terms of EER or minDCF. This observation is in line with the well known fact that VAD is useful in speaker verification. rVAD overall performs better than Sohn *et al.*, Kaldi Energy VAD and is comparable to or marginally better than VQVAD (almost identical in EER and slightly better in minDCF). rVAD-fast shows a similar performance to rVAD in terms of SV, although its VAD performance (as observed in Table 4) is worse than that of rVAD. Considering the huge VAD performance gap (by a factor of more than three in EER) between rVAD and VQVAD as shown in Tables 1 and 5, we conclude that superior performance in VAD does not necessarily translate into an improvement in SV performance. This could be explained by the fact that SV is a task of making one single decision based on the entire sequence/utterance, which differs from the VAD task where each short segment matters. VQVAD is specially optimized for SV.

7.1. Sensitivity of rVAD threshold β on speaker verification performance

Table 10 shows the effect of varying threshold value β in Eq. (8) of rVAD on TD-SV performance. MSNE is used for the second pass denoising and both passes are applied. The results show that the performance of rVAD does not change rapidly with changing the value of β, demonstrating the stability of rVAD. The results, which are obtained but not included in Table 10, also show that rVAD with MSNE-mod performs slightly better than rVAD with MSNE.

Table 10: *Performance of rVAD with various threshold β values for speaker verification on RedDots part-01 (male).*

Methods	Threshold	[%EER/minDCFx100]			Average
	(β)	Target-wrong (TW)	Impostors-correct (IC)	Impostors-wrong (IW)	(%EER/minDCF)
rVAD (MSNE)	0.1	4.08/1.591	2.77/1.243	1.07/0.266	2.64/1.033
	0.2	3.91/1.534	2.80/1.242	1.11/0.278	2.61/**1.018**
	0.3	3.86/1.547	2.82/1.249	1.01/0.275	**2.56**/1.024
	0.4 (default)	3.79/1.572	2.93/1.328	1.01/0.273	2.58/1.058
	0.5	4.36/1.646	3.26/1.451	1.14/0.331	2.92/1.143
	0.6	4.19/1.674	3.20/1.460	1.26/0.325	2.88/1.153
	0.7	4.45/1.735	3.54/1.540	1.29/0.388	3.09/1.221

7.2. Text-dependent speaker verification under noise conditions

We further evaluate the performance of VAD methods for TD-SV under mismatched conditions, where noisy test utterances are scored against the speaker models trained under clean condition (office environment). In order to cover different real-world scenarios, various types of noise are artificially added to the test data with various SNR values. The scaling factor is calculated using the ITU speech voltmeter [76]. TD-SV results of different VAD methods are presented in Table 11. It is observed that TD-SV performances are significantly degraded under noisy conditions as also expected. rVAD achieves mostly lower EER values than those of Kaldi and Sohn et al. and *No-VAD* (i.e., without using VAD) over different noise types and SNR values, but performs slightly worse than VQVAD that is specially designed for SV. Sohn *et al.* VAD provides decent improvement as well. Kaldi Energy VAD degrades the performance compared with No-VAD, as in the case of clean condition.

rVAD-fast gives comparable performance to that of rVAD under noise type of babble, market and car, but it does not work under white noise as the spectral flatness measure is severely affected by white noise. A close analysis shows that for white noise from $0dB$ through $10dB$, spectral flatness values are mostly close to 1.0, as illustrated in Fig.5, due to the similar amount of power in all spectral bands. Therefore, the threshold value of $\theta_{sft} = 0.5$ does not output any speech frames with pitch (and thus no speech frames) for most of noisy test trails in SV and in this experiment, we consider these trails without speech frames as an SV error or misclassification. Numbers of these trails (without any speech frame being detected) are shown in parenthesis in Table 11. When calculating EER, genuine/true trials without speech frames are directly rejected (false rejection) by assigning the lowest score value available in the non-target trials to these genuine trials and vice-versa, namely non-target trails without speech frames are directly accepted (false alarm) by assigning the highest score value available in the genuine trials to these non-target trials.

Table 11: *TD-SV performance of different VAD methods under noisy test environments on RedDots part-01 (male). Numbers in parenthesis show numbers of test utterances (out of total* 3854 *unique utterances) that do not yield any speech frames by the respective VAD methods.*

Noise type	SNR (dB)	% Average EER [TW, IC, IW] across VAD methods									
		no VAD	Kaldi	Sohn et al.	VQ VAD	rVAD			rVAD-fast		
						MSNE (default)	MMSE	mod. MSNE	MSNE	MMSE	mod. MSNE
Clean	-	3.48	3.85	2.73	**2.51**	2.58	2.70	**2.54**	2.63(2)	2.99(2)	**2.52**(2)
White	0	35.23	37.82	37.05(119)	30.18	34.46	34.44	35.29(65)	99.34(3796)	99.38(3796)	99.29(3796)
	05	26.78	30.63	26.68(33)	21.98	25.49	24.82	25.58(11)	68.05(2086)	68.05(2080)	67.84(2079)
	10	18.88	22.19	17.08	14.57	16.42	16.90	17.22	25.26(333)	25.26(326)	25.25(335)
	15	12.47	15.28	10.50	9.28	10.20	10.33	10.41	13.04(61)	13.13(81)	13.30(63)
	20	8.25	9.90	6.59	5.97	6.64	6.60	6.62	6.85(3)	7.12(3)	7.05(3)
Babble	0	35.02	34.42	33.25	30.91	32.88	32.78	33.10	33.01	33.44	33.21
	05	25.17	24.19	22.89	20.78	21.84	22.32	22.36	22.84	23.79	21.90
	10	16.45	15.61	14.17	12.48	13.49	13.72	13.60	14.01(1)	14.43	13.61
	15	10.30	10.26	8.48	7.55	8.19	8.47	8.24	8.37(2)	8.65	8.41
	20	6.34	6.94	5.24	4.66	5.28	5.31	5.17	5.46(2)	5.43(2)	5.59(2)
Market	0	25.39	26.24	24.67	22.79	23.83	24.15	24.09(1)	24.37	24.49	23.94
	05	16.84	17.41	15.51	14.09	15.07	15.33	15.16	15.81	15.38	15.11
	10	10.58	11.11	9.02	8.40	8.84	9.06	9.11	9.23	9.37	9.33
	15	6.58	7.45	5.65	5.06	5.61	5.90	5.73	6.00(1)	6.06(1)	5.97(1)
	20	4.82	5.44	3.90	3.60	3.99	4.10	3.98	4.03(2)	4.21(2)	4.04(2)
Car	0	4.10	5.60	3.70	3.53	3.74	3.95	3.75	3.75	3.74	3.69
	05	3.67	5.01	3.20	3.09	3.23	3.39	3.24	3.24	3.23	3.18
	10	3.33	4.61	3.00	2.90	2.90	3.05	2.96	2.92	2.93	2.88
	15	3.24	4.36	2.80	2.72	2.77	2.85	2.71	2.77	2.78	2.71
	20	3.18	4.18	2.70	2.58	2.62	2.71	2.60	2.63(1)	2.70(1)	**2.56**(1)
Average		13.83	14.93	12.80	11.35	12.37	12.50	12.54	18.54	18.67	18.44

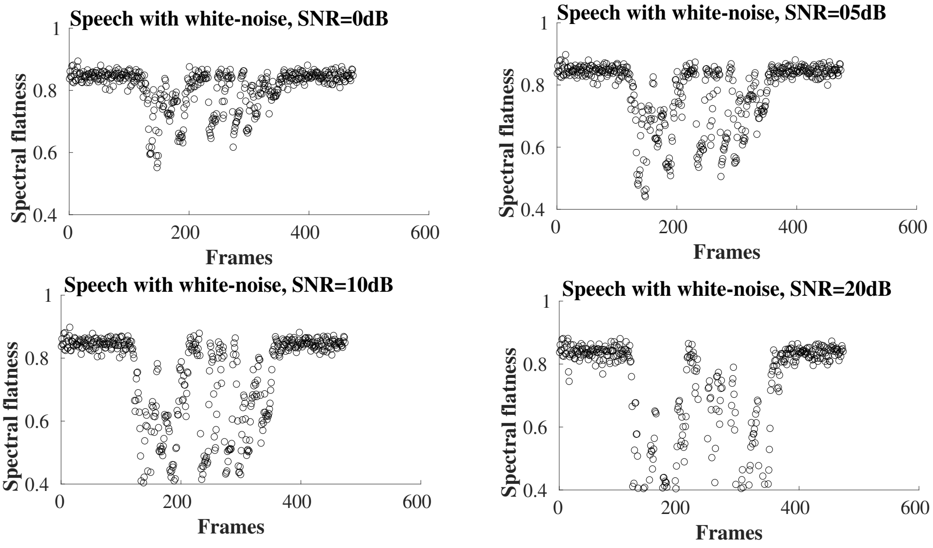


Figure 5: *Scatter plots of SFT values of a speech signal with white-noise for different SNR values.*

7.3. Sensitivity of rVAD threshold β on speaker verification performance under noisy conditions

Table 12 shows the effect of varying threshold β in Eq. (8) of rVAD on the performance of TD-SV under noisy test conditions. It is observed that rVAD is stable towards varying the threshold value.

Table 12: *Performance of rVAD with various threshold β values for speaker verification on RedDots part-01 (male) under noisy conditions.*

Test	SNR	Threshold (β) [% average EER (TW, IC, IW)]								
condition	(dB)	0.1	0.2	0.3	0.4 (default)	0.5	0.6	0.7	0.8	0.9
White	0	34.55	34.37	34.16	34.46	34.41	33.91	**33.47**	34.27	34.76
	05	25.79	25.26	25.09	25.49	25.40	25.28	**24.47**	25.62	25.39
	10	17.40	16.87	16.71	16.42	16.75	16.95	**16.11**	17.27	17.63
	15	10.87	10.66	10.61	10.20	10.42	10.65	**10.10**	11.20	11.69
	20	6.79	6.71	**6.49**	6.64	6.79	6.74	6.64	7.45	7.78
Babble	0	33.79	33.02	33.07	32.88	32.81	32.34	**31.71**	31.63	32.13
	05	23.85	23.45	22.65	21.84	21.86	21.42	**20.84**	21.49	21.95
	10	14.82	14.42	13.88	13.49	13.66	13.15	**12.79**	13.60	13.72
	15	8.92	8.56	8.36	8.19	8.30	7.97	**7.95**	8.37	8.59
	20	5.82	5.54	**5.26**	5.28	5.28	5.30	5.37	5.86	6.28
Market	0	24.43	23.87	23.96	23.86	23.97	**23.08**	23.36	24.09	24.53
	05	15.54	15.22	14.75	15.07	15.18	14.68	**14.36**	15.21	15.55
	10	9.62	9.02	**8.76**	8.84	9.02	9.01	8.83	9.71	9.88
	15	6.21	5.85	5.71	**5.61**	6.02	5.86	5.94	6.61	6.79
	20	4.20	4.10	**3.99**	**3.99**	4.25	4.15	4.51	4.82	5.09
Car	0	3.74	**3.69**	3.76	3.74	4.05	4.05	4.30	5.02	5.21
	05	3.26	**3.21**	3.26	3.23	3.60	3.60	3.78	4.51	4.86
	10	2.94	**2.87**	2.97	2.90	3.30	3.30	3.52	4.17	4.49
	15	**2.71**	2.75	2.72	2.77	3.07	3.09	3.29	3.97	4.14
	20	2.65	2.63	2.63	**2.62**	2.92	2.92	3.14	3.79	4.00
Average		12.89	12.60	12.43	12.37	12.55	12.37	**12.22**	12.93	13.22

8. Conclusion

In this paper, we presented an unsupervised segment-based robust voice activity detection (rVAD) method for voice activity detection (VAD). It consists of two-pass denoising, extended pitch segment detection, and voice activity detection. The first pass denoising uses pitch as a speech indicator to remove high-energy noise segments detected by using *a posteriori* signal-to-noise ratio (SNR) weighted energy difference, while the second pass denoising attempts to remove more stationary noise using speech enhancement methods. Then, extended pitch segments are found. In the end, *a posteriori* SNR weighted energy difference is applied to extended pitch segments of the denoised speech signal for VAD. We evaluated the performance of the proposed rVAD method for VAD and speaker verification (SV) tasks on several diverse databases containing a large variety of noise conditions and compared rVAD against 16 VAD methods including both supervised and unsupervised methods. Experiment results show that the proposed method is compared favourably with a number of existing methods. It is worth to emphasize that rVAD obtained the promising performances across databases and tasks by using the same parameters, indicating the good generalization ability of rVAD. It can be concluded that pitch is a good indicator or anchor for locating speech segments and *a posteriori* signal-to-noise ratio (SNR) weighted energy difference is an effective measure for segmenting speech in noisy environment. Furthermore, a VAD targeted towards SV is able to perform well even though its VAD performance does not.

In addition, we presented a modified version of the rVAD method, called rVAD-fast, where computationally time-consuming pitch extraction is replaced by computationally efficient spectral flatness calculation. The modified version significantly reduces the computational complexity at the cost of moderate inferior VAD performance, which is an advantage when processing a large amount of data and running on resource-limited devices. rVAD-fast, however, breaks down under white noise and therefore it should be used with caution. It has shown to work well under babble, market and car noise and under clean condition. One further finding is that spectral flatness is a good indicator for whether or not there is pitch in a segment as long as the signal is not severely corrupted by white noise.

Overall, it can be concluded that rVAD performs well in both clean and noisy conditions and for both VAD itself and SV. The generalization ability across databases, noisy conditions and tasks was proofed as well.

Future work includes investigating the optimal configurations of rVAD for different applications. The performance of rVAD on automatic speech recognition is worth to study as well.

9. Acknowledgement

Zheng-Hua Tan wishes to thank Dr. James Glass for hosting him in 2012 and 2017 at MIT where the work was done in part.

This work is partly supported by the iSocioBot project, funded by the Danish Council for Independent Research - Technology and Production Sciences (#1335-00162), and the Horizon 2020 OCTAVE Project (#647850), funded by the Research European Agency (REA) of the European Commission.

References

[1] M. Price, J. Glass, A. P. Chandrakasan, A low-power speech recognizer and voice activity detector using deep neural networks, IEEE Journal of Solid-State Circuits 51 (1) (2018) 66–75.

[2] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, P. Dumouche, Supervised/unsupervised voice activity detectors for text-dependent speaker recognition on the rsr2015 corpus, in: Proc. of Odyssey Speaker and Language Recognition Workshop, 2014.

[3] Z.-H. Tan, B. Lindberg, Low-complexity variable frame rate analysis for speech recognition and voice activity detection, IEEE Journal of Selected Topics in Signal Processing 4 (5) (2010) 798 – 807.

[4] J. Ramirez, C. Segura, C. Benitez, A. Torre, A. Rubio, A new kullback-leibler vad for speech recognition in noise, IEEE Signal Processing Letters 11 (2) (2004) 266–269.

[5] X.-L. Zhang, J. Wu, Deep belief networks based voice activity detection, IEEE Transactions on Audio, Speech, and Language Processing 21 (4) (2013) 697–710.

[6] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, V. Mitra, A noise-robust system for nist 2012 speaker recognition evaluation, in: Proc. of Interspeech, 2013, pp. 1981–1985.

[7] L. Ferrer, M. Graciarena, V. Mitra, A phonetically aware system for speech activity detection, in: Proc. of IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP), 2016, pp. 5710–5714.

[8] K. Walker, S. Strassel, The rats radio traffic collection system, in: Proc. of Odyssey Speaker and Language Recognition Workshop, 2012, pp. 291–297.

[9] T. Petsatodis, C. Boukis, F. Talantzis, Z.-H. Tan, R. Prasad, Convex combination of multiple statistical models with application to vad, IEEE Transactions on Audio, Speech, and Language Processing 19 (8) (2011) 2314–2327.

[10] T. Petsatodis, F. Talantzis, C. Boukis, Z.-H. Tan, R. Prasad, Multi-sensor voice activity detection based on multiple observation hypothesis testing, in: Proc. of Interspeech, 2011, pp. 2633–2636.

[11] T. Kinnunen, P. Rajan, A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data, in: Proc. of IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP), 2013.

[12] A. Sholokhov, M. Sahidullah, T. Kinnunen, Semi-supervised speech activity detection with an application to automatic speaker verification, Computer Speech & Language 47 (2018) 132–156.

[13] E. Chuangsuwanich, J. Glass, Robust voice activity detector for real world applications using harmonicity and modulation frequency, in: Proc. of Interspeech, 2011, pp. 2645–264.

[14] S. Sadjadi, J. H. L. Hansen, Unsupervised speech activity detection using voicing measures and perceptual spectral flux, IEEE Signal Processing Letters 20 (3) (2013) 197–200.

[15] T. Kinnunen, A. Sholokhov, E. Khoury, D. A. L. Thomsen, M. Sahidullah, Z.-H. Tan, Happy team entry to nist opensad challenge: A fusion of short-term unsupervised and segment i-vector based speech activity detectors, in: Proc. of Interspeech, 2016, pp. 2992–2996.

[16] A. Drygajlo, Entropy based voice activity detection in very noisy conditions, in: Proc. of EUROSPEECH, 2001, pp. 1887–1890.

[17] D. Vlaj, B. Kotnik, B. Horvat, Z. Kacic, A computationally efficient mel-filter bank vad algorithm for distributed speech recognition systems, EURASIP Journal of Applied Signal Processing 2005 (4) (2005) 487–497.

[18] Z.-H. Tan, B. Lindberg, A posteriori snr weighted energy based variable frame rate analysis for speech recognition, in: Proc. of Interspeech, 2008, pp. 1024–1027.

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

[20] J. F. Bonastre, N. Scheffer, C. Fredouille, D. Matrouf, Nist04 speaker recognition evaluation campaign: new lia speaker detection platform based on alize toolkit, in: Proc. of NIST 2004 speaker recognition workshop, 2004.

[21] E. Khoury, M. Garland, I-vectors for speech activity detection, in: Proc. of Odyssey Speaker and Language Recognition Workshop, 2016, pp. 334–339.

[22] D. L. Hu, et al., Voice activity detection with decision trees in noisy environments, Applied Mechanics and Materials 128-129 (2012) 749–752.

[23] D. Enqing, L. Guizhong, Z. Yatong, Z. Xiaodi, Applying support vector machines to voice activity detection, in: Proc. of International Conference on Spoken Language Processing, 2002.

[24] F. Tao, C. Busso, Bimodal recurrent neural network for audiovisual voice activity detection, in: Proc. of Interspeech, 2017, pp. 1938–1942.

[25] ETSI, Speech processing, transmission and quality aspects (STQ): Distributed speech recognition, advanced front-end feature extraction algorithm, compression algorithm ES 202 050 v1.1.5, ETSI, Geneve, 2007.

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

[27] D. Vlaj, Z. Kai, M. Kos, Voice activity detection algorithm using nonlinear spectral weights, hangover and hang before criteria, Computers & Electrical Engineering 38 (6) (2012) 1820–1836.

[28] ITU, Coding of speech at 8 kbit/s using conjugate structure algebraic code-excited linear-prediction (cs-acelp) annex b: A silence compression scheme, Tech. rep., ITU Recommendation G.729, Geneve (1996).

[29] ITU, Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s. annex a: Silence compression scheme, Tech. rep., ITU Recommendation G.723.1, Geneve (1996).

[30] H.-G. Hirsch, D. Pearce, The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, in: Automatic Speech Recognition: Challenges for the Next Millennium, ISCA ITRW ASR2000, 2000.

[31] Y. Shao, Q. Lin, Use of pitch continuity for robust speech activity detection, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5534–5538.

[32] X.-K. Yang, L. He, D. Qu, W.-Q. Zhang, Voice activity detection algorithm based on long-term pitch information, EURASIP Journal on Audio, Speech, and Music Processing 2016 (1) (2016) 14.

[33] Y. Zhao, Z.-Q. Wang, D. Wang, A two-stage algorithm for noisy and reverberant speech enhancement, in: Proc. of IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP), IEEE, 2017, pp. 5580–5584.

[34] G. Peeters, A large set of audio features for sound description (similarity and classification), Tech. rep., CUIDADO project (2004).

[35] N. Madhu, Note on measures for spectral flatness, Electronics Letters 45 (23) (2009) 1195 – 1196.

[36] M. Moattar, M. M. Homayounpour, A simple but efficient real-time voice activity detection algorithm, in: Proc. of EUSIPCO, 2009, pp. 2549–2553.

[37] https://www.darpa.mil/program/robust-automatic-transcription-of-speech.

[38] The reddots challenge: Towards characterizing speakers from short utterances, https://sites.google.com/site/thereddotsproject/reddots-challenge.

[39] I. Stefanus, R. S. J. Sarwono, M. I. Mandasari, Gmm based automatic speaker verification system development for forensics in bahasa indonesia, in: Proc. of Instrumentation, Control, and Automation (ICA), 2017.

[40] A. Nautsch, R. Bamberger, C. Busch, Decision robustness of voice activity segmentation in unconstrained mobile speaker recognition environments, in: Proc. of Biometrics Special Interest Group (BIOSIG), 2016.

[41] B. K. Dhanush, et al., Factor analysis methods for joint speaker verification and spoof detection, in: Proc. of IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP), 2017.

[42] S. Shepstone, Z.-H. Tan, S. Jensen, Audio-based age and gender identification to enhance the recommendation of tv content, IEEE Transactions on Consumer Electronics 59 (3) (2013) 721–729.

[43] H. Dubey, M. R. Mehl, K. Mankodiya, Bigear: Inferring the ambient and emotional correlates from smartphone-based acoustic big data, in: International Workshop on Big Data Analytics for Smart and Connected, 2016. URL http://arxiv.org/abs/1606.03636

[44] N. Semwal, A. Kumar, S. Narayanan, Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models, in: Identity, Security and Behavior Analysis (ISBA), 2017.

[45] A. Chorianopoulou, P. Koutsakis, A. Potamianos, Speech emotion recognition using affective saliency, in: Proc. of Interspeech, 2016, pp. 500–504.

[46] C.-Y. Lee, Discovering linguistic structures in speech: Models and applications, Ph.D. thesis, Massachusetts Institute of Technology (2014).

[47] Z.-H. Tan, N. B. Thomsen, X. Duan, E. Vlachos, S. E. Shepstone, M. H. Rasmussen, J. L. Hjvang, isociobot: A multimodal interactive social robot, International Journal of Social Robotics (2017) 1–15.

[48] R. Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics, IEEE Transactions on Speech and Audio Processing 9 (5) (2001) 504–512.

[49] P. C. Loizou, Speech enhancement: theory and practice, CRC press, 2007.

[50] Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, IEEE Trans. on Acoust., Speech, Signal Processing 32 (1984) 1109–1121.

[51] T. Gerkman, R. C. Hendriks, Unbiased mmse-based noise power estimation with low complexity and low tracking delay, IEEE Trans Audio, Speech, Language Processing 20 (2012) 1383–1393.

[52] I. Cohen, B. Berdugo, Noise estimation by minima controlled recursive averaging for robust speech enhancement, IEEE Signal Processing Letters 9 (1) (2002) 12–15.

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

[54] J. Jensen, Z.-H. Tan, Minimum mean-square error estimation of mel-frequency cepstral features, IEEE/ACM Trans. Audio, Speech and Language Processing 23 (1) (2015) 186–197.

[55] M. Kolbæk, Z.-H. Tan, J. Jensen, Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems, IEEE/ACM Trans. on Audio, Speech and Language Processing 25 (1) (2017) 153–167.

[56] X. Zhang, Z.-Q. Wang, D. Wang, A speech enhancement algorithm by iterating single- and multi-microphone processing and its application to robust asr, in: Proc. of IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP), 2017, pp. 276–280.

[57] D. Michelsanti, Z.-H. Tan, Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification, in: Proc. of Interspeech, 2017, pp. 2008–2012.

[58] M. Kolbæk, Z.-H. Tan, J. Jensen, Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification, in: Proc. of Spoken Language Technology Workshop (SLT), 2016, pp. 305–311.

[59] J. D. Johnston, Transform coding of audio signals using perceptual noise criteria, IEEE Journal on Selected Areas in Communications 6 (2) (1988) 314–332.

[60] Speaker recognition evaluation 2016, https://www.nist.gov/itl/iad/mig/speaker-recognition-evaluation-2016.

[61] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, V. Zue, TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1, Web Download. Philadelphia: Linguistic Data Consortium, 1993.

[62] A. Larcher, K. A. Lee, B. Ma, H. Li, Text-dependent Speaker Verification: Classifiers, Databases and RSR2015, Speech Communication 60 (2014) 56–77.

[63] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilci, M. Sahidullah, A. Sizov, ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures challenge, in: Proc. of Interspeech, 2015, pp. 2037–2041.

[64] S. Young, D. Kershaw, J. Odell, V. Valtchev, P. Woodland, et al., HTK Book, Copyright 2001-2006 CUED.

[65] I. Kraljevski, Z.-H. Tan, M. P. Bissiri, Comparison of forced-alignment speech recognition and humans for generating reference vad, in: Proc. of Interspeech, 2015, pp. 2937–941.

[66] S. Gonzalez, M. Brookes, A pitch estimation filter robust to high levels of noise (pefac), in: Proc. of EUSIPCO, 2011, pp. 451–455.

[67] P. Boersma, D. Weenink, Praat: doing phonetics by computer (version 5.1.05) [computer program], [Online; Accessed 2009] (2009).
URL `http://www.praat.org/`

[68] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, The kaldi speech recognition toolkit, in: Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding, 2011.

[69] K. Walker, X. Ma, D. Graff, S. Strassel, S. Sessa, K. Jones, Rats speech activity detection, in: LDC2015S02. Hard Drive. Philadelphia: Linguistic Data Consortium, 2015.
URL `https://catalog.ldc.upenn.edu/ldc2015s02`

[70] Evaluation plan for the nist open evaluation of speech activity detection (opensad15) (2015).
URL `https://www.nist.gov/itl/iad/mig/nist-open-speech-activity-detection-evaluation`

[71] C. Kim, R. Stern, Power-normalized cepstral coefficients (pncc) for robust speech recognition, in: Proc. of IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP), 2012, pp. 4101–4104.

[72] O. Plchot, S. Matsoukas, P. Matejka, N. Dehak, J. Z. Ma, S. Cumani, O. Glembek, H. Hermansky, S. H. R. Mallidi, N. Mesgarani, R. M. Schwartz, M. Soufifar, Z. Tan, S. Thomas, B. Zhang, X. Zhou, Developing a speaker identification system for the DARPA RATS project, in: Proc. of IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP), 2013, pp. 6768–6772.

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

[74] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Vesel, P. Matejka, Developing a speech activity detection system for the darpa rats program, in: Proc. of Interspeech, 2012, pp. 1969–1972.

[75] H. Hermanksy, N. Morgan, Rasta processing of speech, IEEE Trans. on Speech and Audio Processing 2 (1994) 578–589.

[76] G. 191, Software tools for speech and audio coding standardization, Tech. rep., International Telecommunication Union (2005).

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: