Digital Signal Processing 141 (2023) 104151

Contents lists available at ScienceDirect

Digital Signal Processing

journal homepage: www.elsevier.com/locate/dsp

A robust and lightweight voice activity detection algorithm for speech enhancement at low signal-to-noise ratio

Zhehui Zhu, Lijun Zhang*, Kaikun Pei, Siqi Chen

School of Automotive Studies, Tongji University, Shanghai 201804, China

ARTICLE INFO

Article history:
Available online 13 July 2023

Keywords:
Voice activity detection
Noise robust
Speech enhancement
Hybrid feature
Machine learning

ABSTRACT

Voice Activity Detection (VAD) is a crucial component of Speech Enhancement (SE) for accurately estimating noise, which directly affects the SE effectiveness in improving speech quality. However, conventional non-data-driven VADs often suffer from decreased accuracy at a low signal-to-noise ratio (SNR). To address this issue, a multi-feature and cosine similarity-based multi-observation VAD algorithm (mVAD) are proposed in this study. This algorithm selects noise-robust features, with Mel-frequency Cepstral Coefficients (MFCCs) as the main features, and utilizes several optimization techniques and an adaptive threshold for background noise updating. Furthermore, the soft VAD results are smoothed with an improved exponential moving average (EMA) algorithm. Besides, a shifting window is utilized to track the mean value and obtain an adaptive threshold for converting the soft results to binary ones. Experimental results indicate that mVAD can maintain high classification accuracy down to -10 dB with an increment of approximately 28% while also being computationally efficient for the CPU time (about 1/3 of statistical model-based methods). It also maintained high robustness at SNRs less than 0 dB ($\boldsymbol{\Delta} \leq 2.1\%$). Moreover, sometimes mVAD even achieved higher accuracy levels than deep learning-based VADs. To further demonstrate the effectiveness of the proposed method, the VAD results are used as an additional feature to train and test a neural network (NN)-based SE model, enhancing the SE performance. This study proves that mVAD does not rely on prior noise knowledge, reaching the dual effect of complexity reduction and accuracy improvement for speech enhancement, making it a promising approach for robust VAD in low SNR environments.

1. Introduction

1.1. Background

Nowadays, speech interaction has gained significant prominence as a widely utilized modality for human-computer interaction, finding applications in intelligent home robots, mobile communications, and, particularly intelligent cockpit systems [1]. However, speech signals employed in in-vehicle applications are susceptible to contamination from external factors that can degrade their quality, including environmental noise, background noise, and reverberation. Consequently, Speech Enhancement (SE) assumes a pivotal role in speech signal processing, aiming to mitigate or suppress additive noise in speech signals, enhancing their clarity and overall quality, as depicted in Fig. 1.

Speech enhancement methods such as spectral subtraction (SS) [1,2] and minimum mean square error-based log-spectral amplitude estimator (MMSE-LSA) [3] rely on accurate knowledge of the background noise. Precise estimation of the noise spectrum amplitude is essential for producing high-quality and intelligible enhanced speech [4]. Nevertheless, conventional methods often lead to speech distortion due to inaccurate noise amplitude estimation. Voice Activity Detection (VAD) can effectively differentiate speech segments from background noise segments, update noise information, and optimize the noise reduction strategy. However, VAD models frequently encounter challenges in extracting speech signals effectively under low Signal-to-Noise Ratio (SNR) conditions. Consequently, enhancing the accuracy and robustness of VAD in low SNR scenarios has gained significant attention. Moreover, with the growing applications of speech technologies in human-robot interaction and mobile communications, there is a demand for VAD algorithms that are lightweight, highly efficient, and capable of meeting low-latency requirements [4].

* Corresponding author.
E-mail address: tjedu_zhanglijun@tongji.edu.cn (L. Zhang).

https://doi.org/10.1016/j.dsp.2023.104151
1051-2004/

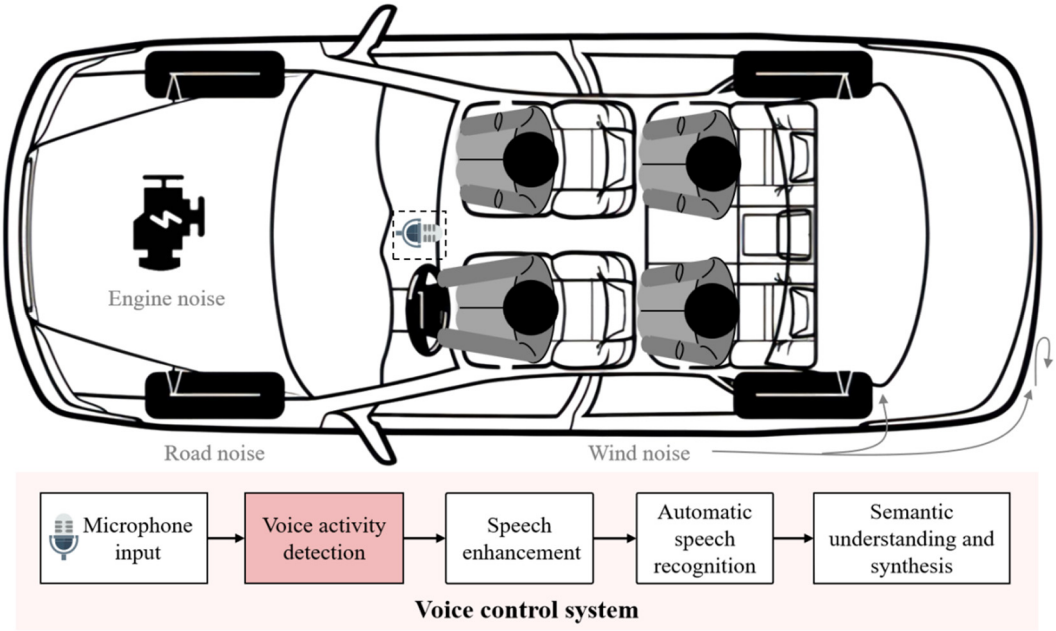

Fig. 1. Schematic diagram of the in-vehicle voice control system.

1.2. Approaches to Voice Activity Detection

VAD methods can be classified into unsupervised and supervised approaches [5]. Unsupervised methods encompass single-feature-based threshold methods, such as energy [6], zero crossing rate [7], pitch [8], and spectral entropy [9]. These methods exhibit satisfactory performance at high SNR but experience substantial degradation at low SNR due to the influence of high-energy noise on speech features. Furthermore, unsupervised methods encompass statistical model-based approaches that depend on assumptions about the distributions of speech and noise, such as Gaussian [10], gamma [10,11], Laplace-Gaussian [12] distribution-based single feature likelihood ratio test (LRT) models and multiple observation likelihood ratio test (MO-LRT) models [13]. However, these methods may encounter difficulties in real-world environments where the underlying assumptions are not met, particularly in low SNR conditions. Tan et al. [14] proposed an unsupervised VAD approach that incorporates two stages of denoising followed by a VAD stage. This method has showcased state-of-the-art performance in noisy environments. The initial pass aims to identify high-energy noise segments, while in the second pass, a speech enhancement (SE) module is employed to denoise the speech corrupted by noise. Despite demonstrating state-of-the-art performance, most of the computational burden in this approach is attributed to pitch detection, resulting in longer CPU processing time than required to achieve the desired efficiency. Although a fast version is proposed to fulfill high-efficiency requirements, it exhibits vulnerabilities when encountering white noise scenarios.

Supervised Voice Activity Detection (VAD) methods have witnessed notable advancements in accuracy through the use of deep learning techniques such as Deep Neural Networks (DNN) [15], Convolutional Neural Networks (CNN) [16], and Long Short Time Memory (LSTM) [17]. However, directly employing deep learning systems for speech and non-speech classification may not fully leverage speech information in the presence of noise, thereby compromising model robustness. One contributing factor to this limitation is the lack of robustness in the network's encoder [16], necessitating individual training for each noise type to achieve satisfactory performance across diverse scenarios [16]. Deep learning-based VAD models have become increasingly intricate and resource-intensive, with over 150 K parameters. Moreover, Attention mechanisms and multi-modal techniques have recently gained traction in enhancing deep learning-based VADs [18–22]. While these approaches can demonstrate state-of-the-art performance by augmenting input features and learning parameters, they also raise implementation concerns on lightweight and remote devices.

VAD can be regarded as a classification algorithm encompassing feature extraction and classification stages. To achieve desirable VAD performance, it is crucial to extract robust features capable of handling diverse noise types and energy levels [23]. Simon et al. [23] conducted a comparative analysis of established VAD features targeting various speech properties, and they found that Mel-frequency cepstral coefficients (MFCC) outperformed others across all SNR ranges from -5 to -15 dB. Consequently, numerous VAD algorithms have been proposed based on MFCC. For instance, Cao et al. [25] introduced a VAD method that utilized a fixed threshold based on MFCC cosine distance, resulting in a 10% improvement in speech recognition rate at 0 dB. However, the lack of adaptivity in the threshold selection rendered the results unsatisfactory when SNR was below 0 dB. Wu et al. [26] employed the first dimension of MFCC and spectral entropy as hybrid feature parameters. They employed clustering and Bayesian information criterion algorithms to estimate an adaptive threshold for VAD, leading to significant accuracy improvement at -5 dB. Muralishankar et al. [27] proposed a modified Mel-DCT-based long-term differential entropy (MMD-LDE) feature for VAD, incorporating an adaptive threshold to enhance accuracy. However, this approach was computationally intensive and exhibited relatively poor real-time performance.

The limitations of the existing methods that are proposed are listed below:

(1) Unsupervised methods, particularly threshold-based VAD, suffer from significant degradation at low SNR. Meantime, statistical model-based VAD often fails to perform effectively in real-world environments when assumptions are not met, especially under low SNR conditions. The computational intensity, and real-time performance are also crucial problems to figure out.

(2) In supervised methods, the direct utilization of deep learning systems sometimes fails to fully exploit speech information in the presence of noise, thereby diminishing model robustness. Moreover, the increased complexity of such methods raises con-

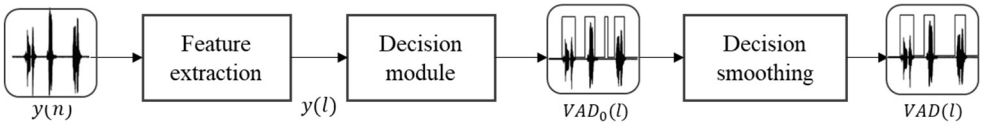

Fig. 2. Block diagram of the VAD in this study.

cerns regarding their implementation on lightweight and remote devices.

To address these limitations, this study proposes a novel multi-feature and multi-observation-based VAD algorithm (mVAD) that leverages hybrid features of different scales and harnesses the benefits of two-step adaptive thresholds with high efficiency. Our work aims to overcome the lack of robustness and complexity associated with existing methods.

1.3. Main work and contribution

The main objective of this study is to propose a robust and lightweight VAD algorithm capable of delivering high-accuracy and low-latency results for speech enhancement modules. Existing research primarily focuses on extracting a single feature for unsupervised methods and employing complex networks for supervised ones. Therefore, ensuring high robustness and efficiency for remote applications in low SNR environments remains challenging. Consequently, this study proposes an mVAD algorithm that utilizes a hybrid feature based on cosine similarity, incorporating multiple features and observations. The feature includes Mel-scale Frequency Cepstral Coefficients (MFCC), Linear Prediction Coefficient (LPC), and Normalized Sub-band Spectral Centroid (NSSC).

This work draws inspiration from recent research [24], which suggests the effectiveness of hybrid features in speech feature extraction. In particular, studies by Zhang et al. [28] found that combining multiple acoustic features can improve VAD's robustness. Wu et al. [29] fed hybrid features, which consist of 2 sub-features weighted by a balancing factor, into support vector machines, and the weighting coefficients were obtained from training. Lei et al. [30] also proposed integrating multiple features. However, their approach relied on manual-dependent thresholds for each parameter, leading to subpar generalization performance in low SNR conditions. mVAD distinguishes itself from previous methods by employing frame-wise multidimensional feature stacking and incorporating multiple observation techniques for MFCC features. The algorithm produces a soft result based on cosine similarity. Binary classification results can be derived utilizing an adaptive threshold that tracks the mean value within a shifting window. Also different from other deep learning methods, mVAD is based on feature extraction and adaptive threshold decision module. The hyperparameters are fixed by preliminary experiments but not through back-propagation, which needs a large number of training data. At the same time, other deep learning methods may combine voice activity detection tasks with speech enhancement in the way of multi-task learning, which brings more complexity. However, in this study, we concatenate mVAD with a speech enhancement network consisting of several layers to evaluate the performance of mVAD. Experimental results demonstrate that mVAD achieves high classification accuracy (Average ACC = 90.2%) at −10 dB and exhibits high robustness ($\boldsymbol{\Delta} \leq 2.1\%$) even when SNR is less than 0 dB.

The major contributions of the proposed work are summarized below:

(1) Proposing a lightweight and robust VAD algorithm (mVAD) for speech enhancement.

(2) Enhancing the robustness of mVAD at low SNR by thoroughly leveraging speech information instead of relying solely on deep learning methods.

(3) mVAD occasionally outperforms deep learning methods when combined with suitable techniques.

(4) Integrating mVAD with a neural network-based speech enhancement (SE) model, with soft VAD results proving more suitable for the front-end processing unit of speech enhancement than binary results.

(5) mVAD holds substantial value in applying lightweight and remote devices, such as in-vehicle voice systems, hearing aids, and intercoms.

This study comprehensively describes the proposed algorithm, covering feature extraction to the decision module. Building upon this, numerous experiments are conducted, encompassing parameter selection, accuracy assessment, and efficiency comparison. Additionally, joint training of a basic speech enhancement model and mVAD is performed to validate the beneficial impact of the proposed method on the speech enhancement module.

2. Proposed VAD method

2.1. Problem formulation

Voice activity detection is a classification problem. The conventional VAD model includes stages of feature extraction, decision module, and decision smoothing [31], as shown in Fig. 2.

Since the speech signal sampled in real application scenarios always contains noise, the task of VAD is to distinguish between frames containing speech and frames only containing noise. It can be assumed that the noise here is additional, so the sampled noisy speech signal $y(n)$ is modeled as the sum of a clean speech signal $x(n)$ and a noise signal $d(n)$:

$$y(n) = x(n) + d(n) \tag{1}$$

As depicted in Fig. 2, feature extraction is primarily conducted at the frame level after obtaining the noisy speech signal. This stage involves pre-emphasis, frame segmentation, and windowing. Speech features for extraction can be classified into time, frequency, and time-frequency domains. Parameters within the frequency domain are typically preferred due to their stability against noise. Finally, the smoothing decision stage is executed by constraining the decision outcome or utilizing a smoothing algorithm to minimize classification error rates [32].

2.2. Feature extraction

The selection of features for extraction in the algorithm is based on preliminary experiments conducted to assess their robustness, as described in Section 4.1. Specifically, the chosen features include Mel-scale Frequency Cepstral Coefficients (MFCCs) with two different observation window lengths, which capture the spectral characteristics of speech. Additionally, Linear Prediction Coefficients (LPC) are employed to represent the time-frequency characteristics of the speech signal. Furthermore, Normalized Sub-band Spectral Centroid (NSSC) is utilized to characterize the tonal properties of the signal.

1) Mel-scale Frequency Cepstral Coefficients (MFCC) [33]: MFCCs are extensively employed in speech recognition due to their ability to represent speech information in the frequency domain accurately. These coefficients are derived from an inverted spectrum

representation obtained by applying the fast Fourier transform (FFT) to short-time windowed signals. Unlike the conventional inverted spectrum representation, MFCCs adopt a nonlinear frequency scale that approximates the acoustic properties of the human ear.

The extraction process of MFCCs involves several steps. Firstly, the voice signal is divided into time frames. Each frame is then subjected to windowing using a Hamming Window function [34]. Next, the FFT is applied to each windowed frame, resulting in the frequency domain representation of the signal. The energy spectrum obtained is subsequently passed through Mel scale filters, which sum the energy across each filter:

$$S(i,m) = \sum_{k=0}^{N-1} E(i,k) H_m(k), \quad 0 \le m \le M \tag{2}$$

where $E(i,k)$ is the energy spectrum of the signal in the frequency domain. The band-pass triangle filter $H_m(k)$ is used to smooth the frequency spectrum, eliminating the effect of harmonics and highlighting the resonance peak of the original speech. The number of filter M is usually set to $22 \sim 26$ [35]. Finally, a discrete cosine transformation (DCT) of the logarithm of the energy is conducted to obtain MFCCs:

$$mfcc(i,r) = \sqrt{\frac{2}{M}} \sum_{m=0}^{M-1} \log[S(i,m)] \cos\left[\frac{\pi r(2m-1)}{2M}\right] \tag{3}$$

where r is the spectral index after DCT, the order of MFCCs is commonly set to $12 \sim 16$ [35].

2) Linear Prediction Coefficient (LPC)**:** The fundamental principle of Linear Predictive Coding (LPC) [36] is a fundamental technique that approximates the sampling values of a speech sample as a linear combination of past sampling values. The derivation of LPC involves minimizing the Mean Square Error (MMSE) and obtaining a unique set of prediction coefficients. LPC is commonly utilized as a feature in speech recognition and enhancement applications [7,37]. The calculation of LPC coefficients can be achieved using the autocorrelation method [38], although this study does not delve into the detailed derivation.

3) Normalized Sub-band Spectral Centroid (NSSC)**:** The Sub-band Spectral Centroid (SSC) is a vital parameter that characterizes the tonal properties of a signal. It represents the center of gravity in Hertz of the frequency distribution [39], analogous to the resonance peak frequency that provides essential information about the energy and frequency distribution of sound signals. Research by Rahul [40] suggests that a VAD based on a weighted spectral centroid outperforms energy-based methods in non-stationary noise. SSC is renowned for its noise robustness and ability to capture the fundamental frequency characteristics of primary harmonic signals. As a result, it is commonly employed in speech enhancement approaches that utilize filtering techniques [39,41].

SSC is solved as follows: assuming the frequency band $[0, F_s/2]$ is divided into M sub-bands, where F_s is the sampling frequency. Set h_m as the upper limit frequency and l_m as the lower of the mth sub-band, respectively. Besides, the window function $w_m(f)$ is used, then the centroid of the mth sub-band spectrum C_m is defined as:

$$C_m = \frac{\int_{l_m}^{h_m} f \omega_m(f) P^{\gamma}(f) df}{\int_{l_m}^{h_m} \omega_m(f) P^{\gamma}(f) df} \tag{4}$$

where $P(f)$ is the power spectrum, f represents the frequency bin, and γ is the constant that controls the dynamic range of the power spectrum. It is proved that when γ is set to 1, the system can achieve good performance, so γ is also set to 1. Then SSC is normalized so that its value is not affected by the spectral sub-band selection [42]. The normalization process is shown as follows:

$$NC_m = C_m - (h_m + l_m)/2(h_m - l_m) \tag{5}$$

where NC_m is the normalized SSC (NSSC).

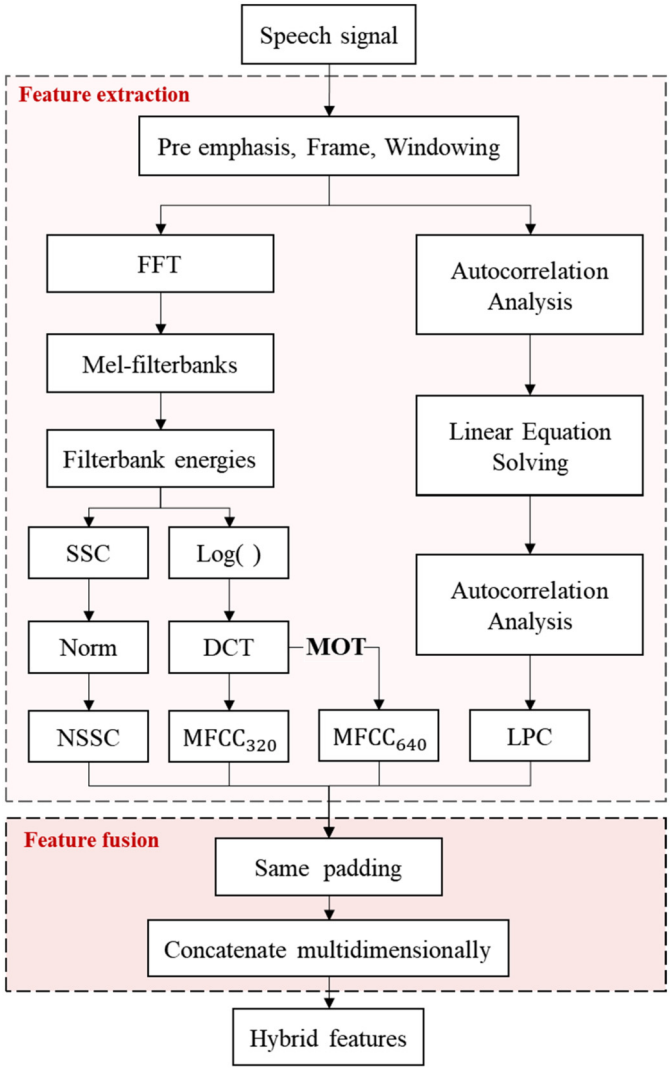

Fig. 3. The procedure of feature extraction and fusion.

2.3. Optimizing stage

The diagram in Fig. 3 illustrates the stages involved in speech feature extraction and fusion. Firstly, the input speech signal $y(n)$ undergoes pre-processing steps, such as pre-emphasis, framing, and windowing. Subsequently, the features are computed at the frame level based on the equations described in Section 4.2. In the fusion stage, the MFCCs are represented as $[m_1, m_2, \ldots, m_k]$, the LPC as $[l_1, l_2, \ldots, l_k]$, and the NSSC as $[n_1, n_2, \ldots, n_k]$, where k denotes the number of speech frames. The values in parentheses following MFCC and NSSC indicate the window length. Finally, the hybrid speech feature $[x_1, x_2, \ldots, x_k]$ is obtained, maintaining the time dimension length equivalent to the number of frames. Specifically, the speech feature for the tth frame is $x_t = [n_t, m_t, l_t]$.

Multiple Observation Technology (MOT) is utilized in the feature extraction stage for the MFCC features obtained. This technique involves extracting speech features using different analysis window lengths, offering two main advantages. Firstly, it enables

the generation of distinct Receiver Operating Characteristic (ROC) curves by treating features with different window lengths as separate entities. Secondly, it effectively suppresses random background noise, enhancing detection efficiency [13]. MOT is specifically applied to MFCCs due to their resilience to various types of noise, leading to improved detection accuracy. As a result, the optimized frame speech feature at frame t is represented by $x_t = [n_{t(320)}, m_{t(320)}, m_{t(640)}, l_t]'$, where the total dimension of MFCCs is 24, comprising 12 from $m_{t(320)}$ and 12 from $m_{t(640)}$. In contrast, the NSSC and LPC dimensions are 1 and 12, respectively. The feature fusion stage differs from methods [28] as it involves multidimensional stacking of features on a frame-by-frame basis and multiple observation techniques for MFCC features.

In the fusion stage, the frame dimensions of the features vary due to the application of the MOT technique, which makes direct concatenation without padding unfeasible. Two padding methods are considered: same padding and linear interpolation padding. The same padding involves duplicating the features of frame t to create frame $t + 1/2$. In contrast, linear interpolation padding utilizes the mean values of the features from frames t and $t + 1$ to generate the features for frames $t + 1/2$. Experimental results indicate that the same padding yields superior outcomes. Consequently, considering the concatenated features, the hybrid feature has a dimension of 37.

2.4. Similarity measurement

Let $x = [x_1, x_2, \ldots, x_k]$ and $y = [y_1, y_2, \ldots, y_k]$ represent two vectors in the feature space, which in the proposed method represents the dimensionally independent hybrid features. The similarity between two vectors can be calculated using distance measurement functions or similarity functions. However, the similarity function is more effective in measuring speech and non-speech features than the distance function [43]. The underlying idea of the proposed method is that the dissimilarity between speech and non-speech segments should be minimized, while the dissimilarity between non-speech and previously estimated noise should be maximized. Therefore, it becomes possible to distinguish between speech and non-speech segments by utilizing a similarity measurement combined with an appropriate decision module.

Commonly used similarity functions for general vectors include the cosine similarity and correlation coefficient. The cosine similarity between vectors x and y is calculated as follows:

$$\cos\alpha = \frac{x_1 y_1 + x_2 y_2 + \ldots + x_k y_k}{\sqrt{(x_1^2 + x_2^2 + \ldots + x_k^2)} \times \sqrt{(y_1^2 + y_2^2 + \ldots + y_k^2)}} \tag{6}$$

The correlation coefficient (Person correlation coefficient) between x and y is:

$$r(x, y) = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{cov}(x, x) \cdot \mathrm{cov}(y, y)}} \tag{7}$$

where $\mathrm{cov}(x, y)$ represents the covariance between x and y; $\mathrm{cov}(x, x)$ and $\mathrm{cov}(y, y)$ represent variance of x and y respectively.

$$\mathrm{cov}(x, y) = \frac{1}{k-1} \sum_{i=1}^{k} (x_i - \overline{x})(y_i - \overline{y}) \tag{8}$$

In the realm of VAD, cosine similarity serves as a valuable tool for evaluating the resemblance between speech and noise components. Conversely, the correlation coefficient is leveraged to estimate the linear correlation between speech and noise signals. To determine the most fitting approach for mVAD, a series of comparative experiments are conducted in Section 4.

2.5. Decision module

The algorithm is based on the assumption that the average of features extracted from the first f_n frames can serve as a representative of the background noise feature. The specific value of f_n is determined according to the method described in the referenced paper [25]. That is:

$$f_n = \frac{0.25 f_s - l_F + l_A}{l_A} \tag{9}$$

where f_s stands for sampling frequency, l_F for frame length, and l_A for advance length. Once the estimated background noise feature is obtained, the similarity between each frame's feature and the background noise feature, as described in Eq. (6) to Eq. (8), is computed. To mitigate the impact of stationary noise and enhance the algorithm's robustness, Cepstral Mean and Variance Normalization (CMVN) is applied.

Furthermore, the background noise frames can be dynamically updated to improve the accuracy of prior background noise estimation. The specific approach involves identifying potential noise frames based on a threshold after the initial distance calculation. Subsequently, the background noise feature is updated. The threshold calculation is performed according to Eq. (10):

$$T_1 = \frac{1}{M} \left(\sum_{i=0}^{M} sorted(x[i]) \right), \quad M = \eta \times nf \tag{10}$$

where the *sorted* function arranges the values in ascending order, and $x[i]$ represents the calculated speech features. The constant M is related to the total number of frames nf. In the experiment, the proportional parameter η is set to 0.15.

Following the update of the background threshold, the algorithm generates a soft VAD result in the form of speech presence probability. To smooth the output, three smoothing techniques are compared: moving average, exponential moving average, and Savitzky-Golay filtering (SG filtering) methods [44]. The moving average method lags significantly behind the current observation, particularly with increasing window size. Conversely, the SG filtering method is suitable for scenarios emphasizing data changes as it effectively preserves signal variation information. Considering real-time requirements and proximity to the current observation, the exponential sliding average method is selected. Furthermore, to address the substantial deviation issue in the initial stage compared to the original method, the improved exponential moving average (iEMA) technique [45] is applied. The calculation of the iEMA can be outlined as follows: (1) Initializing the first smoothed value as the first observed value in the time series. (2) Calculating the smoothed value by taking the weighted average of the current observation and the previous smoothed value by Eq. (11). (3) Repeating step two for all data points in the time series.

$$v_t = [\beta v_{t-1} + (1 - \beta)\theta_t] / \left(1 - \beta^t\right) \tag{11}$$

where θ_t is the presence probability of frame tth speech, and $\beta \in [0, 1)$ is the weighted coefficient. It determines the rate of weights decay, whereby higher values assign greater emphasis to recent observations, while lower values allocate more weight to past observations. Through the iterative application of this formula to each data point, the iEMA method produces a smoothed representation of the data that flexibly adjusts to changes and diminishes the influence of noise and outliers. Significantly, it effectively mitigates the issue of substantial deviation during the initial stage. After the smooth stage, v_t is the final soft voice activity detection result, also represented as v_{soft}.

Table 1
The pseudo-code of mVAD.

Algorithm 1 Multi-feature fusion-based soft and binary VAD

```
Int COS_Sim_VAD (Double x [])
{
  x = Enframe (x []);
  feature_a = concatenate (f_1, f_2, f_3, f_4)
  feature_b = Average (fusion_feature (1: fn)
  Cos [] = CosVec (feature_a, feature_b)
  Cos_n [] = CMVN (1-Cos[])
  For i=fn, . . . , nf
    { if Cos_n [i ]<T1 {
          feature_b.append (Cos_n [i ])
      }
    }
  Cos_updated [] = CosVec (feature_a, feature_b)
  v_soft = EMA (Cos_updated [])
  For j = 0, . . . , nf
    { if V_soft [j]<T2 {
        v_bin [j] = 0
      }
      Else {
          v_bin [j] = 1
      }
    }
  Return v_soft [], v_bin []
}
```

To obtain a binary result, another threshold T_2 judging module that serves to convert the soft results to binary ones is added, as shown in Eq. (12):

$$v_{bin}(i) = \begin{cases} 0 & \text{if } v_{soft}(i) < T_2 \\ 1 & \text{if } v_{soft}(i) \geq T_2 \end{cases} \tag{12}$$

where T_2 is based on the average window tracking method, we set the window length L_w, use the average of the soft results in the window as the threshold, and then make a frame-by-frame judgment, which can be expressed by Eq. (13):

$$T_2 = \frac{1}{K} \sum_{i=k_0}^{k_1} x[i] \tag{13}$$

where k_0 and k_1 are the start and end frames in the kth window, respectively, and K is the number of frames in the kth window.

The flowchart of mVAD is shown in Fig. 4, and the primary process can be expressed as follows:

(1) Calculating the features of all the frames x_a;
(2) With a priori parameters f_n, calculating features x_b of background noise;
(3) Calculating the cosine similarity between the feature of x_a and x_b, and storing the results in a one-dimensional vector *cosVector*[];
(4) Normalizing the result *cosVector*[] by CMVN and then obtaining $cosVector[]_{\text{norm}}$.
(5) Using the calculated threshold T_1 to update the background noise frames and obtain $x_a\_updated$;
(6) Repeating the third step to obtain the one-dimensional vector *cosVector*[] updated and postprocessing the results with the improved EMA to obtain the soft VAD result v_{soft};
(7) Using T_2 to get binary VAD results v_{bin};

The pseudo-code of the algorithm is shown in Table 1.

Considering the time complexity $T(n)$ of the mVAD algorithm, which incorporates the execution of a Time-sort function with a time complexity of $O(nlogn)$, where n represents the length of the noisy speech y, other components of the mVAD algorithm simultaneously operate with a time complexity of $O(n)$. As a result, the overall time complexity of the mVAD algorithm can be determined as $O(nlogn)$. Note that the complexity of [10], [25], and [26] are $O(n^2)$, $O(n)$, $O(n^2)$, respectively.

Based on the comprehensive introduction of the proposed method, the main contributions made by mVAD can be categorized into two aspects:

(1) Feature extraction stage: mVAD introduces hybrid features comprising MFCCs, LPC, and NSSC, distinguishing it from other methods solely relying on MFCC. Additionally, mVAD incorporates complementary features suitable for human voice activity detection and takes advantage of multi-observation techniques in the extraction stage and concatenation feature-wise with a padding strategy in the fusion stage.

(2) Decision stage: mVAD employs a two-step adaptive threshold strategy. The first step involves a prior noise estimation updating module, which gathers potential noise frames to enhance result accuracy. The improved exponential moving average (iEMA) smoothing technique is employed in the second step, followed by applying the second threshold. This approach transforms the soft results into binary ones, a novel utilization of MFCC-based VAD methods.

3. Dataset and evaluation

This section focuses on dataset selection, evaluation metrics, and parameter configuration in mVAD, and introduces several models for comparison.

3.1. Dataset

The dataset preparation for this experimental study follows the methodology outlined in a previous publication [46]. To facilitate hyperparameter tuning experiments, sentences from the TIMIT dev set are utilized, with 10% randomly selected from the train set. Additionally, NOISEX-92 [47] provides various noise types, including babble, two distinct factory noises, oproom, pink, volvo, and white noise. These noises are combined to generate experimental data at −12, −6, and 0 dB, which are crucial for parameter selection in mVAD.

To evaluate mVAD, clean speech data from the TIMIT test dataset is employed. This dataset comprises recordings of 10 male and 10 female speakers, each delivering 20 sentences mixed with a specific type of noise. The noises used include white, babble, and volvo noise, at different signal-to-noise ratio (SNR) levels: 10, 5, 0, −5, and −10 dB. The evaluation process calculates the average result of all test speech samples under a specific SNR condition and noise type, providing a comprehensive assessment of mVAD's performance.

3.2. Evaluation method

ACC and AUC are used as evaluation metrics for comparison with other algorithms. ACC is calculated according to Eq. (12):

$$ACC = \frac{N_{1,1} + N_{0,0}}{nf} \tag{14}$$

where $N_{1,1}$ represents the number of frames correctly classified as speech within the speech category, $N_{0,0}$ represents the number of frames correctly classified as non-speech within the non-speech category, and nf denotes the total number of frames.

In the third sub-experiment, the evaluation metric employed is the Area Under the Curve (AUC) to facilitate comparison with deep learning-based methods. AUC is calculated as the area under the receiver operating characteristic (ROC) curve of the binary classification model. The ROC curve plots the false acceptance rate

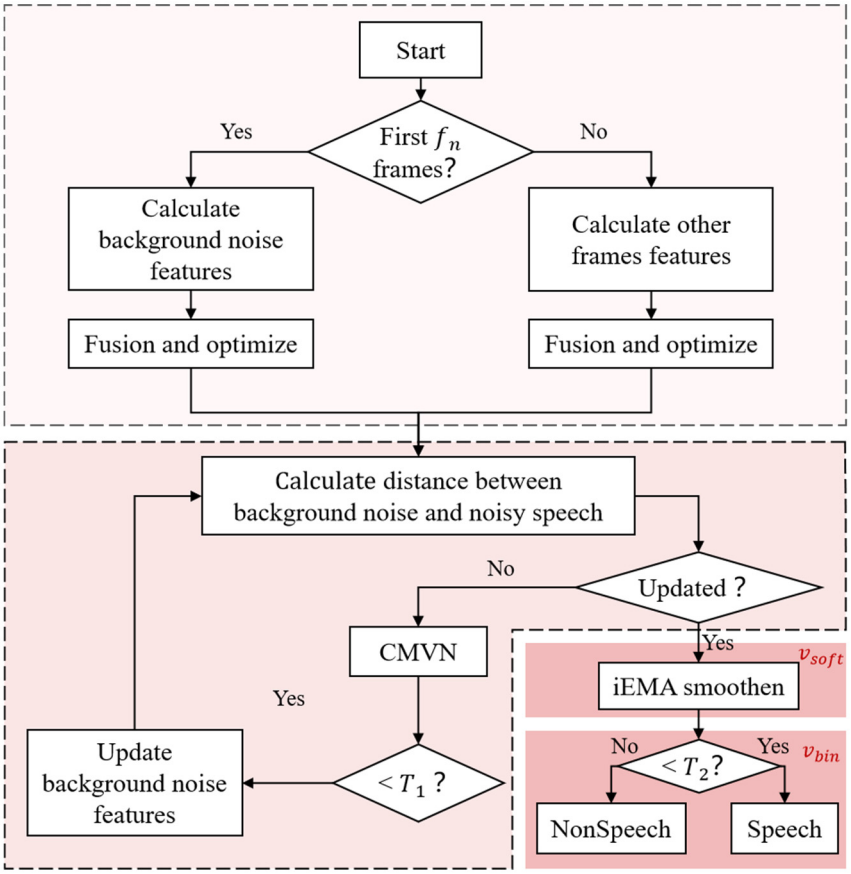

Fig. 4. The flow chart of mVAD.

(FAR) against the false rejection rate (FRR) [48–50]. The AUC metric ranges between 0 and 1, with a higher value indicating better classifier performance [51].

3.3. Methods for comparison

In this section, a concise overview of the algorithms employed for comparative analysis is provided:

(1) Shon's statistical model-based VAD (Shon) [10]: This method employs the decision-directed parameter estimation technique for the likelihood ratio test.

(2) Modified Mel-DCT-based long-term differential entropy-based method (MMD-LDE)[27]: MMD-LDE utilizes a single threshold-based VAD that leverages long-term features. The features are extracted using a modified Mel-discrete cosine transform (MMD) filter bank.

(3) MFCC cos-distance-based method (cos-MFCC) [25]: cos-MFCC relies on a single threshold-based VAD that extracts MFCC cos values.

(4) Product of spectral entropy and Mel-based method (MFPH) [26]: MFPH employs a double threshold-based VAD that incorporates the product of MFCC's first dimensional parameters and spectral entropy. The threshold is adaptively estimated using the fuzzy C-means clustering algorithm (FCM) and Bayesian information criterion (BIC).

(5) rVAD-fast [14]: This method consists of two-pass denoising, extended spectral flatness detection, and voice activity detection. The first pass of denoising removes high-energy noise segments, while the second one aims to eliminate stationary noise using speech enhancement techniques. Then the VAD stage utilizes a posteriori SNR-weighted energy difference.

Table 2
Simulation system settings.

Parameter	Setting
Sampling frequency f_s	16 kHz
Frame length L_F	320 (≡20 ms) | 640 (≡40 ms)
Frame advance L_A	160 (≡10 ms) | 320 (≡20 ms)
FFT length M	1024 (including zero-padding)
Frame overlap	50% (hamming-window)
Threshold T_1 controlling factor η	0.15 (equation (2))
EMA Weighting factor β	0.9
Threshold T_2 tracking win length L_w	6400 (≡0.4 s)

(6) Deep belief networks (DBN) [28]: DBN is a machine learning-based VAD that combines multiple features using DBN to extract a new feature for binary classification.

(7) Generative adversarial network (GAN) [52]: GAN is a deep neural network (DNN)-based VAD consisting of a feature extractor, discriminator, and voice activity detector. The model is trained in a supervised manner.

(8) Masked auditory encoder (AE) based CNN (M-AECNN) [16]: M-AECNN is a DNN-based VAD that employs auditory features extracted by Gammachirp Auditory Filterbank to estimate a masking weight. The masking weight is the input for the classification block to obtain binary output results.

3.4. Parameter settings

Parameters mentioned in the above sections are set according to Table 2.

Table 3
Cosine similarity of noisy speech and clean speech features.

Feature	Dimensionality	Cos similarity
MFCC	12 (Order 2–13)	0.83
LPC	12	0.81
MFCC	16 (Order 2–19)	0.80
MFCC0	1	0.77
NSSC	1	0.77
ZCR	1	0.75
LPS	1	0.68
Spectral entropy	1	0.60
Energy	1	0.54

4. Results and discussion

This section provides a comprehensive analysis of the selection of hyperparameters for extracting the hybrid feature, as listed in Table 2. Subsequently, a series of experiments are conducted to evaluate the performance of the mVAD algorithm. Firstly, the binary results of mVAD are compared with several existing approaches across the full-range SNR range of $[-10, 10]$ dB. To showcase the superior performance of mVAD, a comparison is also made with three machine learning-based methods in a lower SNR range of $[-10, 0]$ dB. Furthermore, the CPU time is calculated to analyze their efficiency. Finally, the integration of mVAD with the front-end processing unit of speech enhancement is explored, and its impact on the speech enhancement (SE) model is examined.

4.1. Parameters of the proposed VAD

1) Hybrid feature selection:

In the algorithm, the features described in Section 2.2 are selected for extraction based on preliminary experiments conducted to assess feature robustness. The cosine similarity between noisy and clean speech is calculated using these features, with a fixed window length 320. Table 3 presents the average experimental results.

The experimental findings demonstrate that frequency-domain parameters exhibit superior noise robustness compared to time-domain parameters. Specifically, MFCCs exhibit more significant resilience compared to energy-based features. However, increasing the dimension of MFCCs does not effectively enhance the test results while it significantly accelerates the computational complexity as the number of parameters increases. Furthermore, extracting only the first dimension of MFCC, known as MFCC0, yields unsatisfactory results. Therefore, the proposed method considers only 12 MFCCs for feature extraction.

Moreover, the zero-crossing rate (ZCR) is found to be highly sensitive to various noise types. Also, at SNR levels below 0 dB, high-energy noise dominates the speech signal, significantly impacting energy-based features, which exhibit lower similarity and less robustness. Based on the experimental results and analysis, the first step of mVAD involves extracting MFCCs, LPC, and NSSC features.

2) Frame length L_F: Frame advance L_A:

The speech features are extracted using different analysis window lengths to incorporate contextual information. We conducted experiments to compare the accuracy of the test set using MFCCs of single window length (i.e., 10) and combined window lengths (i.e., 10&20). The results of these tests are presented in Table 4.

The experimental findings demonstrate that employing longer combined window lengths can effectively mitigate random background noise and enhance the accuracy of mVAD. Additionally, the results indicate that the relationship between these window lengths has a significant impact, with the optimal outcomes observed when they are twice as long. Conversely, adjacent or widely separated window lengths lead to sub-optimal results. Based on these observations, we have chosen to utilize $L_F = 640$ (40 ms) in conjunction with $L_F = 320$ (20 ms) and a 50% overlap.

3) Choice of Threshold T_1 controlling factor η:

The VAD results of noisy and clean sentences are evaluated by calculating the accuracy based on different controlling factors. The relationship between accuracy and the factor η is illustrated in Fig. 5.

The findings indicate that the accuracy tends to decrease as the controlling factor increases. However, there is a peak in accuracy when the factor is in the range of 0.1 to 0.2. Consequently, a value of $\eta = 0.15$ is selected based on this analysis.

4.2. Binary VAD performance ACC comparison

In the experiment, several methods, including Shon [16], MMD-LDE [27], cos-MFCC [25], MFPH [26], and rVAD-fast [14], are selected for comparison, as described in Section 3.3. Additionally, we compare the two similarity functions, correlation coefficient, and cosine similarity, and present the results as mVAD-cor and mVAD-cos. To avoid overfitting, the SNRs and noises used in the following evaluation differ from those used in the parameter selection experiments discussed in Section 4.1. Table 5 illustrates the comparison results in terms of accuracy (ACC).

The experimental results demonstrate that mVAD outperforms the other methods, particularly at low SNR. Notably, the cosine similarity function demonstrates superior accuracy compared to the correlation coefficient across a majority of the conducted tests. When applied in voice activity detection (VAD), cosine similarity exhibits robustness against variations in signal features and places emphasis on the angle between two vectors. On the other hand, the correlation coefficient relies on the assumption of linearity between the signals. This assumption can limit its effectiveness, particularly in situations where the relationship between the speech and noise is nonlinear or complex. As a result, mVAD (referred to as mVAD-cos in the subsequent paragraph) emerges as the preferred choice. Notably, mVAD achieves a remarkable improvement in accuracy of 37% at -10 dB and 25% at -5 dB compared to the statistical model-based methods. In certain scenarios, it even surpasses other algorithms at higher SNRs. For instance, under the Volvo noise scenario, mVAD achieves an ACC of 90.7% at -10 dB, surpassing the ACC of 90.1% at -5 dB in the contrast model.

Additionally, it demonstrates consistent accuracy at low SNRs ($\leq$0 dB) with minimal variation (ΔACC ≤ 0.021), indicating its robustness in low SNR conditions. This improvement can be attributed to the effective fusion of multiple high-robust features. It can also be attributed to the utilization of CMVN to mitigate the impact of stationary noise. Moreover, an improved smoothing technique is incorporated to reduce initial errors and time delays and make the results less susceptible to high-energy noise.

Regarding specific noise types, mVAD-cos achieves higher accuracy in Volvo noise scenarios than other noise types. This observation is supported by the spectrogram of the noise shown in Fig. 6, confirming that the energy in Volvo noise is predominantly concentrated in the low-frequency range. In contrast, babble noise exhibits energy primarily in the high-frequency range, while white noise energy is evenly distributed across the frequency spectrum. As a result, mVAD demonstrates relative robustness to low-frequency noise, with a slight decrease in accuracy observed in babble noise environments. Overall, mVAD-cos exhibits the highest accuracy among the compared methods.

4.3. Soft VAD performance AUC comparison

In the experiment, three machine learning-based methods are selected, DBN [14], GAN [52], and M-AECNN [16]. These methods

Table 4
Test accuracy (%) with different window length.

Win/ms	10	20	30	40	10&20	10&30	10&40	20&30	20&40	30&40
average	83.23	85.74	85.29	85.65	83.62	83.80	83.14	83.56	**87.46**	84.13

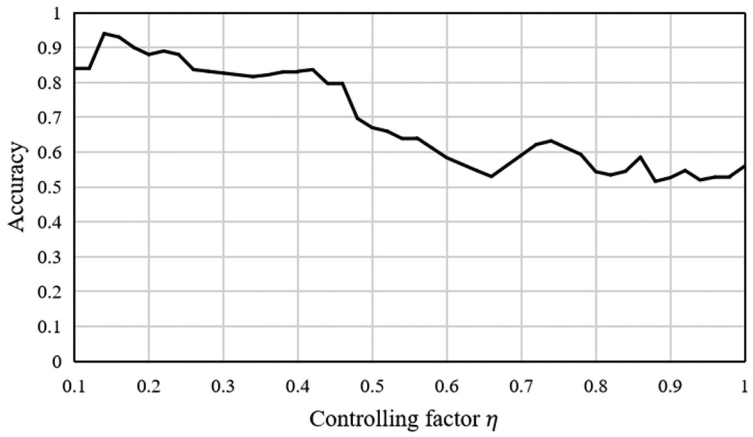

Fig. 5. The relationship between VAD accuracy and Threshold T_1 controlling factor η.

Table 5
Detailed ACC (%) for Test 2.

Noise	SNR	Sohn [10]	MMD-LDE [10]	cos-MFCC [25]	MFPH [25]	rVAD-fast [14]	mVAD-cor	mVAD-cos
White	−10	52.9	83.4	66.1	/	70.1	87.7	**89.8**
	−5	57.3	88.7	70.2	**92.3**	74.5	88.9	90.5
	0	60.0	90.2	78.6	91.1	77.9	**91.2**	91.1
	5	73.3	91.4	80.0	**93.6**	82.2	91.9	92.2
	10	87.4	/	88.5	94.3	85.6	94.0	**96.1**
Babble	−10	50.1	75.1	64.0	/	81.5	85.1	**90.0**
	−5	58.5	82.3	69.2	90.2	85.6	88.9	**91.7**
	0	64.0	87.3	72.1	**92.3**	87.0	90.2	92.1
	5	71.8	89.2	80.0	93.2	90.1	91.2	**93.4**
	10	78.0	/	84.9	**93.9**	92.4	92.7	93.6
Volvo	−10	50.8	79.9	79.0	/	82.1	87.3	**90.1**
	−5	58.8	84.9	82.1	90.1	84.0	89.3	**92.3**
	0	65.0	87.0	85.7	90.5	88.2	90.5	**92.8**
	5	72.4	88.3	86.9	91.1	90.4	91.6	**93.9**
	10	79.4	/	89.4	92.4	93.8	92.8	**94.9**

Table 6
Detailed AUC (%) for Test 3.

Noise	SNR	DBN [28]	GAN [52]*	M-AECNN [16]*	mVAD-cor	mVAD-cos
White	−10	/	86.9	87.1	88.2	**90.2**
	−5	/	**94.4**	92.7	94.1	94.0
	0	/	**97.0**	94.6	96.1	95.0
Babble	−10	**90.2**	86.9	87.1	87.4	89.1
	−5	92.3	**94.4**	92.7	90.2	92.1
	0	93.6	**97.0**	94.6	93.7	96.3
Volvo	−10	91.0	86.9	87.1	90.2	**91.4**
	−5	92.6	**94.4**	92.7	91.2	92.5
	0	93.7	**97.0**	94.6	94.0	94.6

(*represents that they are average results that are tested throughout all types of noises).

were chosen based on their high performance in recent years. Additionally, the M-AECNN was selected due to its smaller parameter size and higher computational efficiency, aligning with the requirements for lightweight VAD applications. These methods were evaluated using the Area Under the Curve (AUC) metric, and the results are presented in Table 6. Furthermore, we plotted the Receiver Operating Characteristic (ROC) curves for Test 3 with SNRs of 0, −5, and −10 dB to assess the detection performance under different noise types and levels. The ROC curves are displayed in Fig. 7.

4.4. Results visualization

The experimental results demonstrate that mVAD achieves comparable accuracy to existing deep learning-based methods, particularly in low SNR conditions. Interestingly, mVAD exhibits slight superiority over these methods in white and volvo noise scenarios at −10 dB. These findings validate the primary objective of mVAD, which aims to provide optimal accuracy while maintaining efficiency and a lightweight algorithm. The experimental results

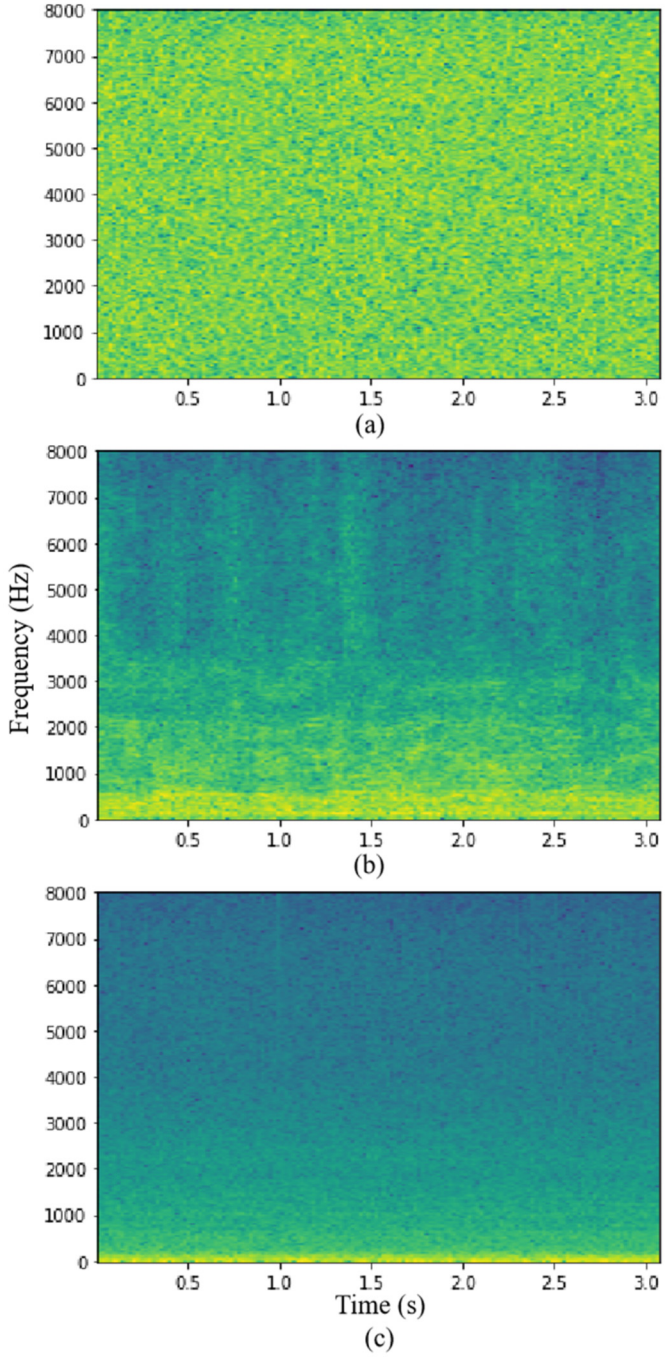

Fig. 6. Spectrogram of three different noises (a. White noise; b. Babble noise; c. Volvo noise).

strongly support the successful attainment of this objective. To visually present the results and facilitate comparison, a test speech sample is selected to showcase the performance of the VAD algorithm under three different noise types, as depicted in Fig. 8. The initial row in the picture illustrates the pristine speech waveform accompanied by its corresponding VAD label, while the subsequent row depicts the speech waveform corrupted by ambient noise. The amplitudes of both waveforms have been normalized within the range of $[-1, 1]$. The following row exhibits the output of the soft VAD, while the final row illustrates the binary VAD result. Notably, both results have been normalized to $[0, 1]$. Fig. 8 (a), (b), and (c) present the mVAD outcomes at a SNR of 0 dB under white noise, babble noise, and volvo noise, respectively.

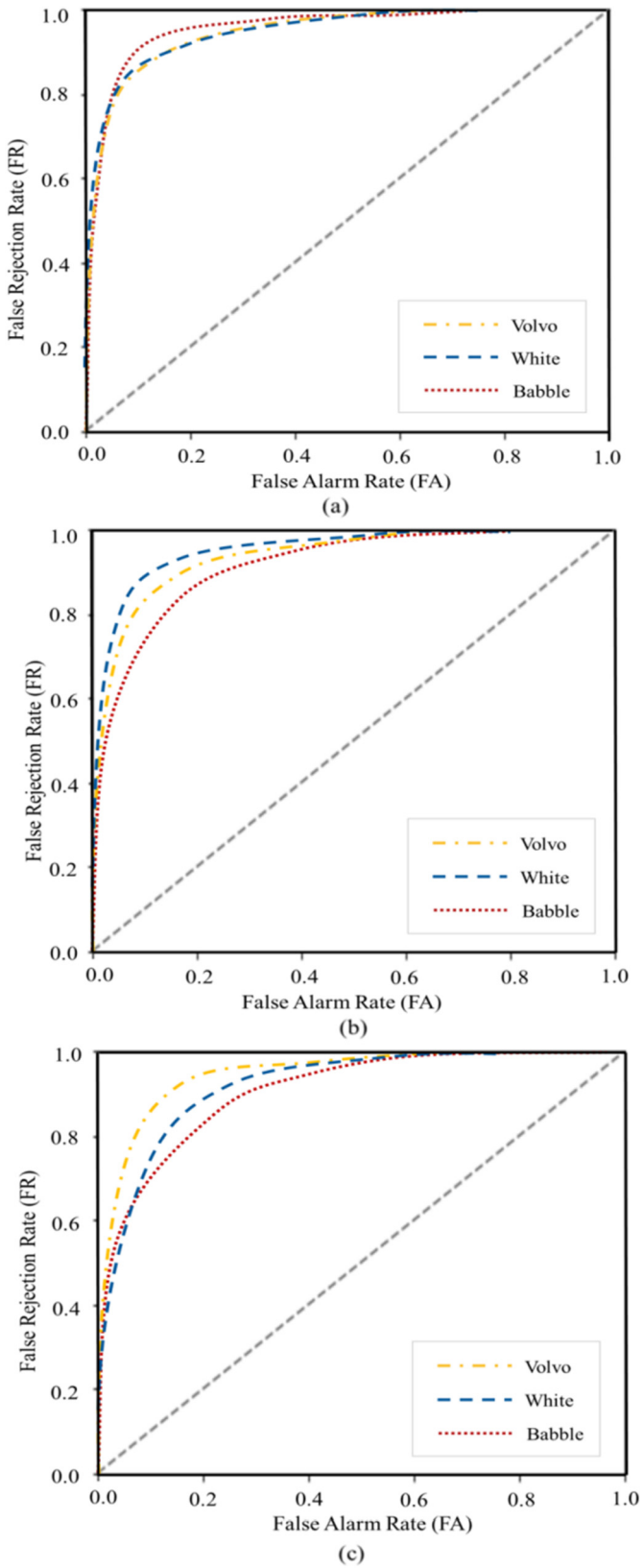

Fig. 7. Probability of false alarm detection versus probability. Acoustic environment: Volvo, White, and Babble noise (a. −10 dB; b. −5 dB; c. 0 dB).

The presented figure reveals that the binary VAD results demonstrate a high accuracy (approximately 98%) in the initial half of the extended speech segment affected by volvo noise. However, the VAD algorithm tends to misclassify speech as background

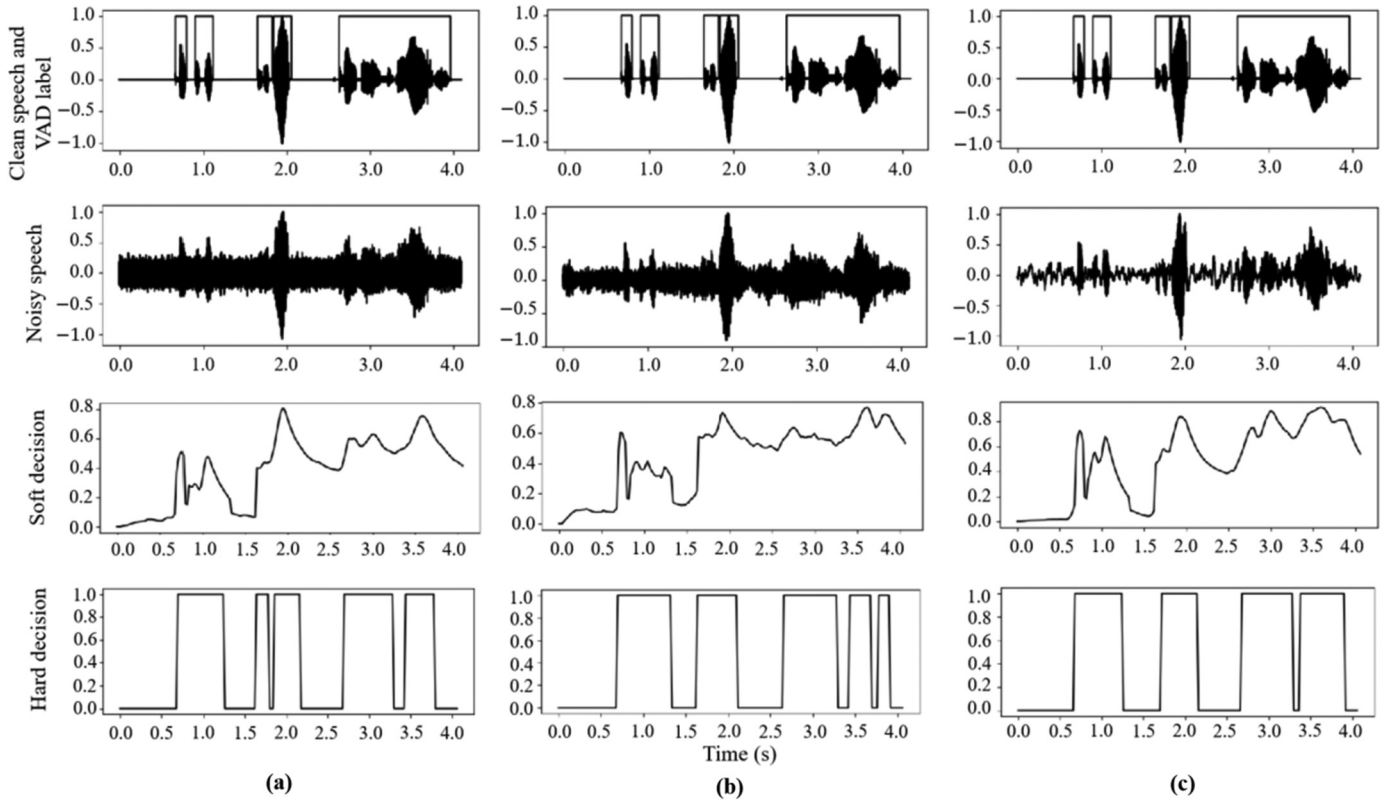

Fig. 8. Results of mVAD with 0 dB SNR under three types of noise (a. White noise; b. Babble noise; c. Volvo noise).

noise in the latter half of the segment, where the energy is relatively lower. This misclassification pattern is also observed in the binary results under the other two noise types, indicating a consistent tendency of mVAD to disregard specific speech segments. Consequently, when utilized in conjunction with speech enhancement and recognition systems, the binary results may fail to detect crucial speech segments.

In contrast, the corresponding soft VAD results exhibit a more pronounced probability of voice presence in the latter half of the segment. It is important to note that this misclassification arises from the binary conversion process, which reduces complexity but at the expense of accuracy. Compared to the binary results, the soft results offer greater flexibility and may prove more beneficial when integrated with speech enhancement models.

4.5. Efficiency comparison

mVAD has been developed for applications in remote communication, such as voice interaction in intelligent cockpits and hearing aids. Therefore, it has to be lightweight. The selection of hyperparameters in the algorithm is based on preliminary experiments rather than a backpropagation process. Moreover, combining hybrid features and an adaptive two-step threshold contributes to the efficiency and accuracy of mVAD without relying on large-scale architectures. To demonstrate the efficiency of mVAD, the CPU time is compared with several other methods in this section. The test is conducted on an Intel i7 CPU, using a total audio length of 6240.29 s. The methods chosen for comparison include rVAD-fast[14], the statistical model-based method [10], cos-MFCC [25], MFPH [26], and the light CNN-based model [26] mentioned in Section 3.3. To reduce the measurement error, five repeated measurements are performed for each test method, and the average results, along with their standard deviations, are presented in Table 7.

The evaluation metric employed during the testing phase is the Real-time working ratio, which quantifies the ratio between CPU processing time and the duration of the input audio. This metric is commonly referred to as the real-time factor (RTF). Additionally, Table 7 presents the number of parameters that necessitate learning and the corresponding time complexities associated with each algorithm. It is worth noting that deep learning-based approaches demand substantial training time to enhance their performance.

According to the results presented in the table, the computational time of mVAD is comparable to that of the MFCC-based method, albeit slightly longer due to the integration of multiple optimization strategies. Compared to rVAD-fast, the proposed method is faster. Furthermore, the computational time is approximately one-third of the statistical model-based and MFPH-based methods, with a remarkable real-time factor of only 0.0209. It is worth noting that while machine learning-based methods may require significant training time, they offer greater computational efficiency compared to traditional methods. Overall, these findings highlight the significant application potential of mVAD in devices that require low latency and high real-time processing capabilities. Examples include in-vehicle speech interaction front-end processing units and hearing aids.

4.6. Speech enhancement combination

Drawing inspiration from conventional speech enhancement methods that utilize VAD for noise spectrum estimation and noise information updating, the mVAD model is seamlessly integrated with a simple DNN-based Speech Enhancement model (NN-SE), as illustrated in Fig. 9.

A five-layer DNN is employed as a baseline for the mapping-speech enhancement training. The training dataset consists of the TIMIT training set and seven types of noises extracted from NOISEX92, including babble, two factory noises, oproom, pink, volvo, and white noise. The noises are mixed at various SNRs (−12, −6,

Table 7
Average CPU time (seconds).

Noise type	mVAD	rVAD-fast	Statistical model-based method	MFCC-based method	MFPH-based method	CNN-based method
params	/	/	/	/	/	1.6 K
O(n)	nlogn	n^2	n^2	n	n^2	n
Total /s	130.352 ±0.008	327.615 ±0.003	407.121 ±0.027	108.954 ±0.012	472.110 ±0.006	52.549 ±0.023
RTF	0.0209	0.0525	0.0652	0.0174	0.0756	0.0084

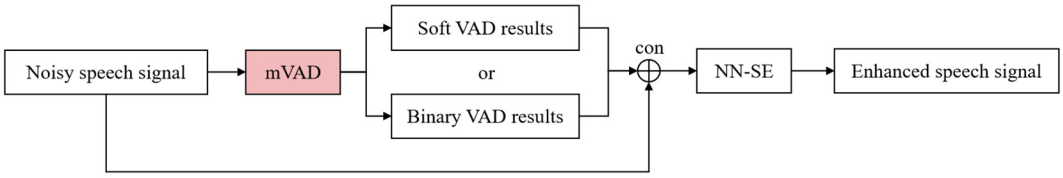

Fig. 9. The combination of mVAD and NN-SE.

Table 8
SE models dataset.

DNN-based SE model	Dataset
Baseline model M1	Noisy speech
M2	Noisy speech + Oracle soft VAD
M3	Noisy speech + Noisy binary VAD
M4	Noisy speech + Noisy soft VAD

Table 9
Testing results.

SNR	M1		M2		M3		M4	
	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
10	1.56	0.79	1.54	0.79	1.56	0.80	1.65	0.81
5	1.42	0.74	1.42	0.76	1.47	0.77	1.51	0.78
0	1.35	0.69	1.37	0.71	1.42	0.74	1.43	0.75
−5	1.28	0.61	1.33	0.68	1.34	0.66	1.35	0.67
−10	1.18	0.55	1.21	0.59	1.25	0.60	1.26	0.58

0, 6, and 12), and all models are trained for 5 epochs. The dataset used for model comparison is outlined in Table 8. The training data is divided into an 80% training set and a 20% validation set. A TIMIT test set mixed with babble, white, and volvo noises at different SNRs (−10, −5, 0, 5, and 10) is utilized to evaluate the model with PESQ and STOI as evaluation metrics.

Table 8 describes four models with different training sets. In M2, the VAD results of clean speech concatenated with noisy speech are used as model input, which represents oracle information not obtainable in real-life scenarios. The difference between M2 and M3 lies in concatenating soft VAD and binary results, respectively. In M4, only the VAD result of the noisy input, which reflects real scenario applications, is considered. The testing results for comparison are presented in Table 9.

The results indicate satisfactory performance for all models when the input speech is relatively clean, with an SNR greater than or equal to 5 dB. However, as the SNR decreases below 0 dB, the models without speech presence possibly exhibit a significant decline in performance, as exemplified by M1 at −10 dB. Therefore, combining VAD results with the speech enhancement (SE) model proves beneficial. Comparing M3 and M4 suggests that soft VAD results align better with the SE model in real-world scenarios. Notably, the soft VAD results for noisy speech yield a more substantial increase in PESQ than for clean speech. However, this discrepancy may stem from a mismatch between the clean VAD training data and the noisy VAD test data. Additionally, Fig. 10 presents a visualization of a speech segment from a speaker in the test dataset enhanced by the four models.

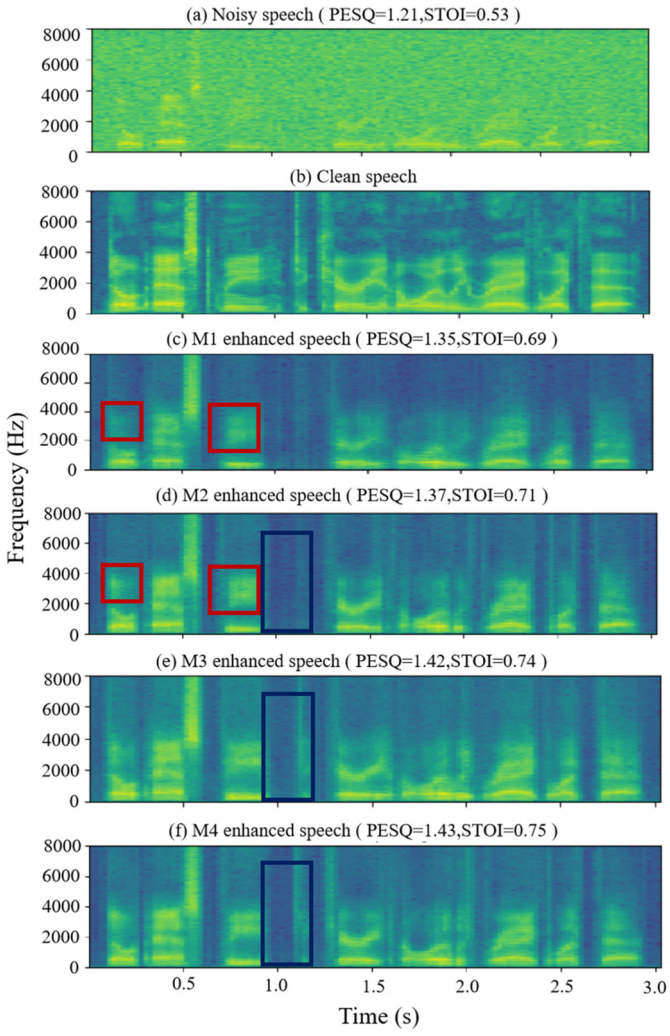

Fig. 10. Spectrograms of an utterance corrupted by white noise at 0 dB and enhanced by 4 models (a. Clean speech; b. Noisy speech; c. M1 enhanced speech; d. M2 enhanced speech; e. M3 enhanced speech; f. M4 enhanced speech).

Fig. 10 showcases the spectrograms of a sample utterance corrupted by white noise at 0 dB and the corresponding enhanced results from four models. In Fig. 10 (b), the red rectangle indicates that the VAD results can assist the SE model in reconstructing more detailed and accurate information for M2-enhanced speech.

However, it should be noted that SE with oracle soft VAD results does not guarantee complete enhanced speech. The model may tend to remove speech information, especially when it is short and overshadowed by noise, as observed in the blue rectangles in Fig. 10 (d), (e), and (f). On the other hand, a model trained with noisy VAD results demonstrates better noise removal and preservation of speech information, as it is trained with a dataset that better represents real-world applications.

Based on the above analysis, it can be inferred that soft VAD results are more accurate and suitable for speech enhancement front-end processing units, surpassing other models, particularly at low SNRs. For example, at 0 dB of white noise added to the speech, there is an improvement of 0.08 PESQ (from 1.35 for M1 to 1.43 for M4) and 0.06 STOI (from 0.69 for M1 to 0.75 for M4). It is essential to mention that a simple four-layer DNN network is employed in this study to evaluate the significance of VAD information for SE models, with training conducted for only 5 epochs. Consequently, achieving the same performance as models with complex structures and high-performance layers is not feasible. Nevertheless, the results demonstrate the critical role of VAD with high robustness at low SNR levels for an SE model.

5. Conclusions

Voice Activity Detection (VAD) plays a crucial role in speech systems, but striking a balance between high accuracy and low complexity poses a challenge. To address this issue and cater to low Signal-to-Noise Ratio (SNR) applications, this paper proposes a VAD method using a feature fusion strategy with multiple observation technology (mVAD). Experimental results demonstrate that mVAD outperforms state-of-the-art traditional methods, maintaining robust performance even at SNRs below 0 dB and achieving comparable or better performance than deep learning-based methods. Notably, mVAD exhibits a CPU processing time of only about one-third of that of statistical model-based methods, with a Real-Time Factor (RTF) as low as 0.0209, making it suitable for real-time performance applications.

Moreover, an adaptive threshold is employed, utilizing a window mean tracking method to convert soft VAD results into binary ones. However, experimental results indicate that the soft results offer greater flexibility and exhibit fewer errors. This is further corroborated by combining the VAD results with a Deep Neural Network (DNN)-based model, revealing the superior accuracy of soft VAD results for speech enhancement front-end processing units. Considering its high accuracy and low complexity, this method is well-suited for remote devices such as voice control systems in intelligent cockpits, wearable devices, and hearing aids. In addition to evaluating performance on simulated data, mVAD holds promise for real-world applications, including integration into vehicle systems through embedded development.

Future research directions include exploring decision schemes in the proposed method and combining them with appropriate NN-based networks. Furthermore, mVAD can be integrated with different high-performance Speech Enhancement (SE) models to enhance the quality and intelligibility of enhanced speech. Ultimately, mVAD can be practically embedded into intelligent cockpits, contributing to improved speech processing and communication experiences.

CRediT authorship contribution statement

Zhehui Zhu: Methodology, Experiments, Conceptualization, Software, Writing – original draft. **Lijun Zhang:** Supervision, Resources, Funding acquisition. **Kaikun Pei:** Data curation, Methodology. **Siqi Chen:** Supervision, Methodology, Writing – review & editing.

Declaration of competing interest

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.

Data availability

Data will be made available on request.

Acknowledgment

The authors appreciate the kind support from the National Natural Science Foundation of China (No. 52072268).

References

[1] S. Wu, et al., Design and implementation of intelligent car controlled by voice, in: 2022 International Conference on Computer Network, Electronic and Automation (ICCNEA), 2022.
[2] H. Gustafsson, S.E. Nordholm, I. Claesson, Spectral subtraction using reduced delay convolution and adaptive averaging, IEEE Trans. Speech Audio Process. 9 (8) (2001) 799–807.
[3] Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator, IEEE Trans. Acoust. Speech Signal Process. 33 (2) (1985) 443–445.
[4] P.C. Loizou, Speech Enhancement: Theory and Practice, CRC Press, 2007.
[5] N.T. Anh, et al., LIS-Net: an end-to-end light interior search network for speech command recognition, Comput. Speech Lang. 65 (2021) 101131.
[6] P. Jing, Spectrum energy based voice activity detection, in: 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), 2017.
[7] A. Benyassine, et al., ITU-T recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications, IEEE Commun. Mag. 35 (9) (1997) 64–73.
[8] M. Thiagarajan, J. Natarajan, K.M. Sharavanaraju, Pitch-based voice activity detection for feedback cancellation and noise reduction in hearing aids, Circuits Syst. Signal Process. 37 (10) (2018) 4504–4526.
[9] K. Wang, Y. Tasi, Voice Activity Detection Algorithm with Low Signal-to-Noise Ratios Based on Spectrum Entropy, IEEE, 2008.
[10] S. Jongseo, S.K. Nam, S. Wonyong, A statistical model-based voice activity detection, IEEE Signal Process. Lett. 6 (1) (1999) 1–3.
[11] R. Tahmasbi, S. Rezaei, A soft voice activity detection using GARCH filter and variance gamma distribution, IEEE Trans. Audio Speech Lang. Process. 15 (4) (2007) 1129–1134.
[12] S. Gazor, W. Zhang, A soft voice activity detector based on a Laplacian-Gaussian model, IEEE Trans. Speech Audio Process. 11 (5) (2003) 498–505.
[13] J. Ramirez, et al., Statistical voice activity detection using a multiple observation likelihood ratio test, IEEE Signal Process. Lett. 12 (10) (2005) 689–692.
[14] Z. Tan, A.K. Sarkar, N. Dehak, rVAD: an unsupervised segment-based robust voice activity detection method, Comput. Speech Lang. 59 (2020) 1–21.
[15] T.G. Kang, N.S. Kim, DNN-based voice activity detection with multi-task learning, IEICE Trans. Inf. Syst. 99 (2) (2016) 550–553.
[16] N. Li, et al., Robust voice activity detection using a masked auditory encoder based convolutional neural network, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.
[17] N. Wilkinson, T. Niesler, A Hybrid CNN-BiLSTM Voice Activity Detector, IEEE, 2021.
[18] I. Hwang, H. Park, J. Chang, Ensemble of deep neural networks using acoustic environment classification for statistical model-based voice activity detection, Comput. Speech Lang. 38 (2016) 1–12.
[19] H. Zhou, et al., Audio-visual information fusion using cross-modal teacher-student learning for voice activity detection in realistic environments, in: Interspeech, 2021.
[20] Y. Hou, et al., Attention-based cross-modal fusion for audio-visual voice activity detection in musical video streams, arXiv preprint, arXiv:2106.11411, 2021.
[21] Y. Hou, et al., Rule-Embedded Network for Audio-Visual Voice Activity Detection in Live Musical Video Streams, IEEE, 2021.
[22] S. Li, et al., Voice activity detection using a local-global attention model, Appl. Acoust. 195 (2022) 108802.
[23] J.A. Haigh, J.S. Mason, Robust voice activity detection using cepstral features, in: Proceedings of TENCON'93. IEEE Region 10 International Conference on Computers, Communications and Automation, 1993.
[24] S. Graf, et al., Features for voice activity detection: a comparative analysis, EURASIP J. Adv. Signal Process. 2015 (1) (2015) 91.
[25] D. Cao, X. Gao, L. Gao, An improved endpoint detection algorithm based on MFCC cosine value, Wirel. Pers. Commun. 95 (3) (2017) 2073–2090.

[26] W.U. Xin-Zhong, et al., Voice activity detection method based on MFPH, J. Beijing Univ. Posts Telecomm. (2019).
[27] R. Muralishankar, D. Ghosh, S. Gurugopinath, A novel modified mel-DCT filter bank structure with application to voice activity detection, IEEE Signal Process. Lett. 27 (2020) 1240–1244.
[28] X.-L. Zhang, J. Wu, Deep belief networks based voice activity detection, IEEE Trans. Audio Speech Lang. Process. 21 (4) (2013) 697–710.
[29] Z. Xiaolei, W. Ji, L. Ping, Support Vector Machine Based VAD Using the Multiple Observation Compound Feature, J. Tsinghua Univ. (Sci. Technol.) 51 (09) (2011) 1209–1214.
[30] L. Jing, H. Peiyu, X. Zili, Adaptive speech endpoint detection based on multi-parameter fusion in low SNR situation, J. Signal Process. 36 (08) (2020) 1205–1211.
[31] J. Ramirez, J.M. Gorriz, J.C. Segura, Voice Activity Detection. Fundamentals and Speech Recognition System Robustness, Robust Speech Recognition and Understanding, 2007.
[32] S.S. Meduri, R. Ananth, A Survey and Evaluation of Voice Activity Detection Algorithms, 2012.
[33] F. Zheng, G. Zhang, Z. Song, Comparison of different implementations of MFCC, J. Comput. Sci. Technol. 16 (6) (2001) 582–589.
[34] P. Podder, et al., Comparative performance analysis of Hamming, hanning and blackman window, Int. J. Comput. Appl. 96 (18) (2014) 1–7.
[35] J.D. Hoyt, H. Wechsler, Detection of human speech in structured noise, in: Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing, 1994.
[36] N. Dave, Feature extraction methods LPC, PLP and MFCC in speech recognition, Int. J. Adv. Res. Eng. Technol. 1 (6) (2013) 1–4.
[37] L. Rabiner, M. Sambur, Application of an LPC distance measure to the voiced-unvoiced-silence detection problem, IEEE Trans. Acoust. Speech Signal Process. 25 (4) (1977) 338–343.
[38] L. Rabiner, M. Sambur, Voiced-unvoiced-silence detection using the Itakura LPC distance measure, in: ICASSP '77. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1977.
[39] K.K. Paliwal, Spectral subband centroids as features for speech recognition, in: 1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings, 1997.
[40] S. Jiasong, Z. Jingyun, Y. Yi, Effective audio fingerprint retrieval based on the spectral sub-band centroid feature, J. Tsinghua Univ. (Sci. Technol.) 57 (04) (2017) 382–387.
[41] R. Ishaq, et al., Subband modulator Kalman filtering for single channel speech enhancement, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.
[42] J.S. Seo, et al., Audio fingerprinting based on normalized spectral subband centroids, in: Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005, 2005.
[43] W. Hongzhi, X. Yuchao, L. Meijing, Study on the MFCC similarity-based voice activity detection algorithm, in: 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011.
[44] W.H. Press, S.A. Teukolsky, Savitzky-Golay smoothing filters, Comput. Phys. 4 (6) (1990) 669–672.
[45] A. Haq, et al., Improved exponentially weighted moving average control charts for monitoring process mean and dispersion, Qual. Reliab. Eng. Int. 31 (2) (2015) 217–237.
[46] F. Heese, M. Niermann, P. Vary, Speech-codebook based soft voice activity detection, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.
[47] J.S. Garofolo, et al., DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM, NIST speech disc 1-1.1. NASA STI/Recon technical report n, 1993, 93: p. 27403.
[48] M. Unoki, et al., Voice activity detection in MTF-based power envelope restoration, in: INTERSPEECH, 2011.
[49] N. Wilkinson, T. Niesler, A hybrid CNN-BiLSTM voice activity detector, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.
[50] Y. Jung, et al., Joint learning using denoising variational autoencoders for voice activity detection, in: INTERSPEECH, 2018.
[51] Z.X. Joint, Training ResCNN-based voice activity detection with speech enhancement, in: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019.
[52] H.K.J. ADA-VAD, Unpaired adversarial domain adaptation for noise-robust voice activity detection, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.

Zhehui Zhu received the B.S. degree in Vehicle engineering from China Agricultural University, Beijing, China, in 2021. She is currently pursuing the M.S. degree with the Department of Automotive Studies, Tongji University, Shanghai, China. Her research interests include Speech signal processing and Speech enhancement.

Lijun Zhang received the Ph.D. degree in Vehicle engineering from Tongji University, Shanghai, China, in 2005. He is currently working as a dean, professor with the Department of Automotive Studies, Tongji University, Shanghai, China. His research interests include Automotive vibration-noise and Intelligent control, Vehicle system dynamics and intelligent control, Integration and control of electric vehicle powertrain, Intelligent vehicle ergonomics and intelligent cockpit.

Kaikun Pei is currently pursuing the Ph.D. degree in Tongji University. His main research areas are Speech signal processing, Speech enhancement and Speaker verification.

Siqi Chen is currently pursuing the Ph.D. degree with the Department of Automotive Studies, Tongji University, Shanghai, China. His research interests include multi physics modeling, AI algorithm application and Optimization for vehicles, Heat and mass transfer analysis.