Applied Acoustics 160 (2020) 107133

Contents lists available at ScienceDirect

Applied Acoustics

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

An overview of speech endpoint detection algorithms

Tao Zhang*, Yangyang Shao, Yaqin Wu, Yanzhang Geng, Long Fan

School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

ARTICLE INFO

Article history:
Received 25 August 2019
Received in revised form 21 October 2019
Accepted 5 November 2019
Available online 21 November 2019

Keywords:
Speech endpoint detection
Time domain
Frequency domain
Cepstrum domain
Neural network

ABSTRACT

Speech endpoint detection is an important part of modern speech information processing technology. The success of endpoint detection directly improves the performance and quality of speech coding, speech recognition, speech synthesis and human interaction. The robustness and detection accuracy of algorithms have always been hot topics for many scholars in the condition of low Signal-to-Noise Ratio (SNR) and complex noise. In this paper, we aim to provide an overview of the state-of-the-art in time domain, frequency domain and cepstrum domain for speech endpoint detection algorithms and to cast a glance at the challenges for future research.

1. Introduction

As an essential tool in our daily communication, language is an external logic expression of the thoughts formed in human brain. As the most natural information carrier, Speech is an acoustic expression of language. Speech endpoint detection is not only the key of speech analysis, speech coding and speech synthesis [1], but also an important part of Voice Internet Protocol, speaker recognition and hand-held telephone [2].

Speech endpoint detection, as its name suggests, aims to determine the starting and ending points of the speech signal. Typically, speech endpoint detection is adapted to the speech front-end processing. For instance, in speech recognition, digital speech signal is composed of voice, silence and various background noise. A key problem is the accuracy of speech endpoint detection, because effective detection technology can not only reduce the system processing time thus realizing the real-time processing of system, but also eliminate the noise interference of silent segments thus improving the recognition performance of subsequent process. Similarly, endpoint detection system [3] plays a vital role in other various speech applications such as isolated word, continuous speech and speaker recognition systems.

An overall speech endpoint detection system mainly consists of four parts including preprocessing, feature extraction, decision execution and endpoint marking.

An ideal speech endpoint detection algorithm needs to meet the following three conditions: 1) Results of the endpoint detection algorithm must be accurate, consistent and reliable; 2) Be able to adapt to background noise, especially in wireless phone; 3) Good real-time performance and low computational complexity are required. The input of speech endpoint detection is noisy signal which is mixed by clean speech and noise in real life. In practice, speech detection algorithms need to meet constraints of various conditions: 1) The energy of speech is obviously higher than that of noise; 2) Noise is stable during a long duration; 3) Speech signal is periodic; 4) The spectrum of speech is more orderly than that of noise, which is due to the fact that noise lacks a specific structure. In the process of speech endpoint detection, extracting appropriate features in the light of various input signals is one problem to be solved. After that, proper decision method is required to detect and mark the endpoints of speech. At present, there are two difficulties in endpoint detection: 1) In low SNR or non-stationary noise conditions, the detection performance is poor; 2) Some of the signals beginning with unvoiced speech, fricatives and bursts are easily drowned by noise, resulting in false detection [4].

The rest of the paper is organized as follows: Section 2 gives an overview of the background of speech endpoint detection. Section 3 reviews the extracted features in various domains. Section 4 reviews the decision methods. Section 5 presents the simulation results of different classical algorithms, followed by the Conclusion and Prospect in Section 6.

2. Background

From the earliest application of speech endpoint detection to date, endpoint detection has undergone a series of developments, as shown in Fig. 1.

* Corresponding author.
E-mail address: zhangtao@tju.edu.cn (T. Zhang).

https://doi.org/10.1016/j.apacoust.2019.107133
0003-682X/© 2019 Elsevier Ltd. All rights reserved.

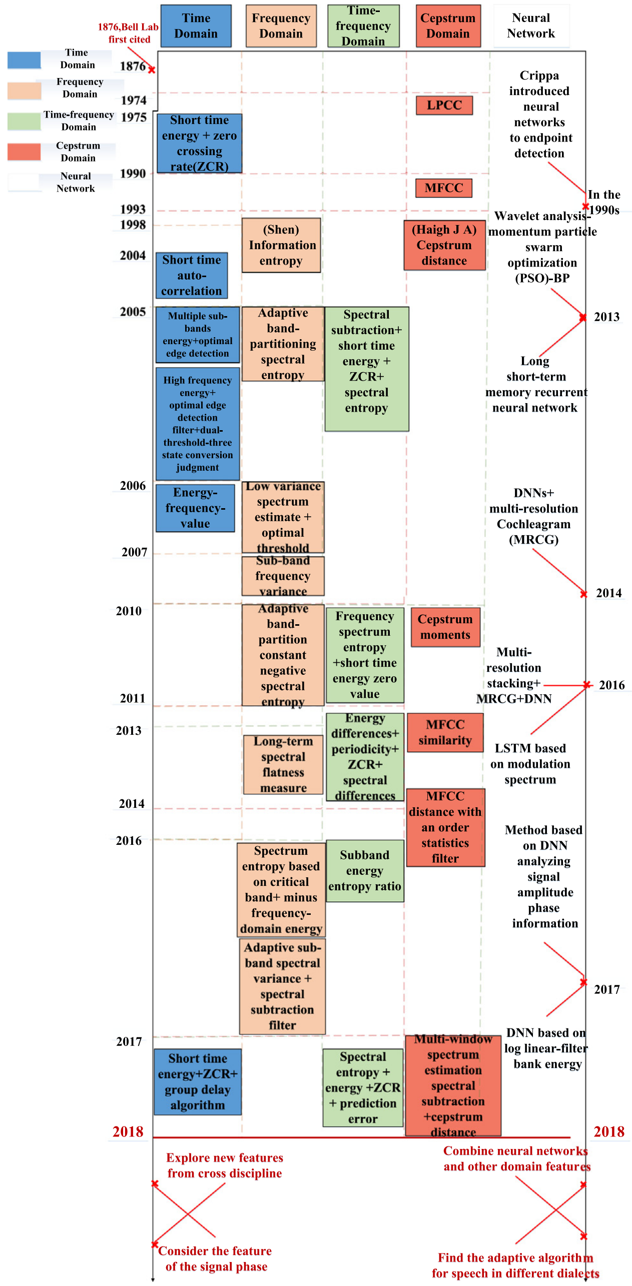

Time Domain
Frequency Domain
Time-frequency Domain
Cepstrum Domain
Neural Network
1876,Bell Lab first cited
1876
1974
1975
1990
1993
1998
2004
2005
2006
2007
2010
2011
2013
2014
2016
2017
2018
LPCC
Short time energy + zero crossing rate(ZCR)
MFCC
(Shen) Information entropy
(Haigh J A) Cepstrum distance
Short time auto-correlation
Multiple sub-bands energy+optimal edge detection
High frequency energy+ optimal edge detection filter+dual-threshold-three state conversion judgment
Adaptive band-partitioning spectral entropy
Spectral subtraction+ short time energy + ZCR+ spectral entropy
Energy-frequency-value
Low variance spectrum estimate + optimal threshold
Sub-band frequency variance
Adaptive band-partition constant negative spectral entropy
Frequency spectrum entropy +short time energy zero value
Cepstrum moments
Long-term spectral flatness measure
Energy differences+ periodicity+ ZCR+ spectral differences
MFCC similarity
MFCC distance with an order statistics filter
Spectrum entropy based on critical band+ minus frequency-domain energy
Subband energy entropy ratio
Adaptive sub-band spectral variance + spectral subtraction filter
Short time energy+ZCR+ group delay algorithm
Spectral entropy + energy +ZCR + prediction error
Multi-window spectrum estimation spectral subtraction +cepstrum distance
Crippa introduced neural networks to endpoint detection
In the 1990s
Wavelet analysis-momentum particle swarm optimization (PSO)-BP
2013
Long short-term memory recurrent neural network
DNNs+ multi-resolution Cochleagram (MRCG)
2014
Multi-resolution stacking+ MRCG+DNN
2016
LSTM based on modulation spectrum
Method based on DNN analyzing signal amplitude phase information
2017
DNN based on log linear-filter bank energy
2018
Explore new features from cross discipline
Consider the feature of the signal phase
Combine neural networks and other domain features
Find the adaptive algorithm for speech in different dialects

Fig. 1. Overview of the development of speech endpoint detection technology.

In Fig. 1, the abscissa is the classification of the domain, the ordinate is the evolution of time. The panes in different colors represent different methods. Here the blue, orange, green and red panes are representations of the speech endpoint methods based on features in time domain, frequency domain, time-frequency domain and cepstrum domain, respectively. While the white panes represent the methods based on neural networks. Fig. 1 is summarized in accordance with time successively on speech endpoint detection algorithms based on different domain characteristics. Finally, a brief forecast for future research directions are listed in Fig. 1. Note that the development speed of neural network keeps different pace with the other four domains.

The first application of speech endpoint detection was in 1876, when Bell Labs introduced speech endpoint detection technology into telephone system. Thus this algorithm realized the redistribution of idle channel and improved the channel utilization of circuit exchange. Following this, a variety of endpoint detection algorithms continued to emerge. The first proposed method was based on the features in time domain. In 1975, Rabiner L. R. used the features combined by short-time energy and Zero Crossing Rate (ZCR) to determine whether a signal was speech or background noise [5]. Many scholars have made a lot of improvements on the algorithms based on the features in time domain. For example, Zulfiqar *et al.* [6] introduced a long-term feature by calculating fractal dimension using Katz algorithm [7] into speech endpoint detection. To some extent, the detection accuracy and the virtual detection rate were improved.

In order to make up for the shortcomings of the features in time domain, the endpoint detection algorithms based on features in frequency domain were presented. In the 1980s, Fast Fourier Transform (FFT) was proposed to directly extract the frequency domain information of speech signals by a Japanese scholar. Subsequently, various features in frequency domain have been emerging, such as the spectral variance, spectral entropy and long-term signal features such as Long-term Signal Variability (LTSV) [8] feature, Long-term Spectral Flatness Measure (LSFM) [9] feature and Long-term Spectral Variability Measure (LSVM) [10] feature. With a series of features emerging both in time domain and frequency domain, scholars have been starting considering whether it is possible to find a better detection algorithm while combining the features both in time domain and frequency domain. On the basis of FFT computation, the features in cepstrum domain have been presented such as Linear Prediction Cepstrum Coefficient (LPCC) [11] and Mel Frequency Cepstral Coefficients (MFCC) [12]. The LPCC feature contains the signal spectrum envelope information, which can be used to estimate the cepstrum of speech signals. Auditory psychophysics shows that the perception of frequency for human ear is non-linear. Based on that, MFCC feature has been presented. In addition to the above three domains, there are other transform decomposition algorithms such as wavelet transform [13] and Teager energy operator algorithm et al. [14], which will be introduced in detail in Section 3.5.

Moreover, neural network is another mainstream method for speech endpoint detection. In the early 1990s, Crippa of Pittsburgh University creatively applied the Artificial Neural Network (ANN) for speech endpoint detection. By making full use of the fast convergence characteristics of feed-forward neural networks, the weights matrix of network could be trained quickly as long as a set of characteristics, thus achieving high detection performance. In order to overcoming some trivial problems and inaccuracy provoked by setting threshold, Chomorlig [15] applies C-SVM to endpoint detection for Mongolian speech. A Wavelet Analysis-Improved Momentum Particle Swarm Optimization-BP (WA-IMPSO-BP) algorithm was applied to perform endpoint detection in Mu Li et al. [16], where the extracted features were used as the input of Back Propagation (BP) neural network optimized by particle swarm algorithm. After that, deep learning approaches have shown high performance for speech endpoint detection. However, such approaches have very long inference times creating hindrance in their utilization in a real-time frame-based speech processing pipeline. To overcome this limitation, [17] referred to a smartphone app that performed the real-time speech endpoint detection in real-time with low audio latency based on Convolutional Neural Network (CNN) [18]. Obviously, this field has been attracting more and more attention.

3. Feature extraction

Feature extraction is a crucial step in speech endpoint detection, which directly determines the performance of executing decision. With the deepening of research for speech signals, more and more features are proposed which yield good detection results.

3.1. Features in time domain

Generally speaking, the methods based on features in time domain are first proposed. The commonly used noises in the early work are all stationary noises at high SNRs. But few research is about noise at low SNRs or complex non-stationary noise conditions. In practical applications, features in time domain and frequency domain are often combined, which can achieve improved endpoint detection results in high SNR environments.

3.1.1. Concepts

The generic endpoint detection algorithms in time domain include short-time energy [19], short-time ZCR [20] and short-time autocorrelation [21]. Short-time energy is the square of the average amplitude of signal. The short-time energy of $y_i(n)$ in the *i-th* frame of speech signal is defined as:

$$E(i) = \sum_{n=0}^{L-1} y_i(n)^2 \tag{1}$$

Among noise, voiced speech and unvoiced speech, noise has the lowest short-time energy, voiced speech has the highest and the unvoiced speech is in the middle. On the basis of this fact, we can distinguish the voice segments and the noise segments. Sometimes short-time energy can be replaced by average amplitude, whose dynamic range is small. In addition, average amplitude is only a simple product weighted summation, resulting in a lower discrimination degree than short-time energy in practical application. Under high SNR conditions, short-time ZCR can be used to distinguish the unvoiced and voiced speeches. The short-time average ZCR represents the times a speech frame crosses the horizontal axis. The short-time average ZCR is defined as:

$$Z(i) = 1/2 \sum_{n=0}^{L-1} |\text{sgn}[y_i(n)] - \text{sgn}[y_i(n-1)]|, 1 \le i \le f_n \tag{2}$$

where L denotes the frame size, $y_i(n)$ denotes the *i-th* frame of the speech signal after windowing and framing, f_n is the total number of frames, and sgn is defined as:

$$\text{sgn}[x] = \begin{cases} 1, x \ge 0 \\ -1, x < 0 \end{cases} \tag{3}$$

The most classical algorithm in time domain is dual-threshold decision method based on ZCR and short-time energy [5]. On these grounds, some scholars constructed the energy-zero ratio feature by calculating the ratio of energy to ZCR. It has been proved that short-time autocorrelation is also an effective speech detection feature in [22] because of the correlation of noise and speech signals. The short-time autocorrelation function is given in Eq. (4).

$$R_i(k) = \sum_{n=0}^{L-k-1} y_i(n) y_i(n+k) \tag{4}$$

where L is the length of the framing signal and k is the delay.

3.1.2. Simulation implementation

In this paper, the tested audio signals in wav format with 8 kHz sampling rate and 16 bits PCM are recorded in a quiet environment. The frame size is 32 ms, namely 256 samples, and the frame shift is 16 ms, which is 128 samples. Here we choose one speech signal to carry out the simulation implementation. The content of the speech signal is: “Fang fa de xuan ze, jue ding le ni de ban shi xiao lv”(In Chinese Pinyin). In the experimental results, each black block represents a detected speech segment.

Fig. 2 shows the result of the dual-threshold decision method based on ZCR and short-time energy. We can see that the algorithm based on short-time energy can detect the high-energy vowels accurately in a pure speech signal. Furthermore, the initial syllables in Chinese Pinyin are usually initial consonants with higher frequencies, and their corresponding ZCRs are large. Therefore, short-time average ZCR can make up for short-time energy to find the starting points. The red lines in Fig. 2 represent the updated starting points of the short-time average ZCR method. However, this algorithm has a large error for single word detection, and cannot accurately detect the endpoints of the syllables such as ‘f’, ‘d’ and ‘x’. Besides, this algorithm cannot effectively handle the pauses between the syllables. For signals mixed with White Gaussian Noise (WGN), the detection accuracy is greatly reduced. In addition, the detection accuracy of algorithm is directly affected by the two threshold values.

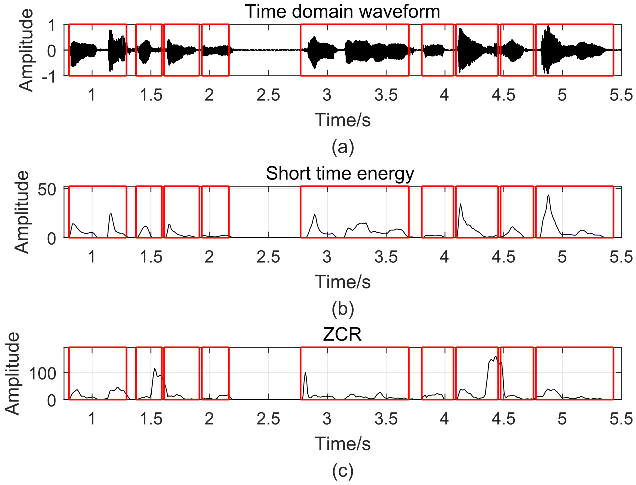


Fig. 2. (a) Waveform of speech “Fang fa de xuan ze, jue ding le ni de ban shi xiao lv”; (b) Speech endpoint detection result using dual-threshold decision method based on and short-time energy; (c) Speech endpoint detection result using dual-threshold decision method based on zero crossing rate. The red blocks are representations of the detected speech segments. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.1.3. Sub-conclusion

Amongst the pioneering studies, many scholars have improved some algorithms based on features in time domain.

Inevitably, a simple dual-threshold decision method could not effectively detect the endpoints, so [23] used short-time high frequency energy as an auxiliary feature, and used the optimal edge detection filter [24] and reasonable dual-threshold-three state conversion judgment [25] at the same time to detect the endpoints on the basis of short-time full-band energy. The decision mechanism of this algorithm in [23] could maintain good performance and robustness for absolute amplitude variations under WGN environments. However, the thresholds of the dual-threshold decision method vary in different noise environments under different SNRs. In order to solve the problem of dynamic adjustment of the thresholds in different SNRs, an algorithm was proposed in [26] by combining the multiple sub-bands energies and optimal edge detection [27] as the decision criteria. The detection accuracies in [26] can be improved under the conditions of white noise, factory noise and vehicle interior noise. In addition to several common noise types studied above, it is necessary to further improve the detection accuracy for the sudden non-stationary noise. In order to adapt the threshold to the strength of background noise automatically, an improved method of endpoint detection based on the transient Energy-Frequency-Value (the product of energy and frequency) was introduced in [28], which effectively enhanced the accuracy of endpoints detecting and its anti-noise performance as mentioned in [28].

Traditional algorithms based on features in time domain are not ideal for specific signal types, so some researchers continue to make improvements. Considering the Maithili speech signal as research object, a new algorithm was proposed in [29], firstly the word boundaries were detected from the voice signals of sentences using short-time energy and ZCR [30], then the rough syllable boundaries were detected by the group delay algorithm. To make these syllable boundaries more accurate, group delay algorithm was modified in [29] by considering the differences of consecutive peaks in the negative derivative phase. The experiment using the data set of 25 different sentences from 10 different speakers was performed and an accuracy rate of 85.62% was achieved using this algorithm in [29].

3.2. Features in frequency domain

The endpoint detection algorithms based on features in time domain are first proposed. Good results have been obtained in the laboratory environment. However in noise environments, the performance of this algorithm is greatly degraded, which is due to the sensitiveness of ZCR to additive noise. In order to get a better result, people start to pay attention to the features in frequency domain. Two classical algorithms, the spectral entropy method and the variance method, are mainly introduced below.

3.2.1. Concepts

Entropy in information theory is a measure of the uncertainty of a random event. When a system is more orderly, its entropy is lower. Conversely, the more chaotic the system is, the higher the entropy is. In 1998, Shen et al. [31] was the first to introduce the concept of information entropy to speech endpoint detection. Because feature is only correlative with the randomness of energy and is independent of energy amplitude of signal, it can avoid a large amount of calculation and be robust to the noise. Here, short-time spectral entropy of each speech frame is defined as:

$$H_i = -\sum_{k=0}^{N/2} p_i(k)\log p_i(k) \tag{5}$$

For noise, the distribution of the normalized spectral probability density function is more uniform and the spectral entropy is larger than that of the speech signal. However, for speech signal, spectral entropy is lower than that of noise because of the spectrum characteristic of formant, so speech signal and noise signal can be detected by this characteristic. However, in a condition of noise with high energy, the performance becomes poor. Hence, an improved algorithm based on the spectral entropy feature was presented in [32]. This algorithm introduced a positive constant K when calculating spectral probability density function of entropy. The introduction of K enhanced the discriminability between speech signals and noise signals and improved the robustness of the spectral entropy so that it became easier to set thresholds. The experimental results in [32] reveal the validity of the improved entropy and prove that the improved entropy outperforms the basic entropy. As shown in [32], compared with an energy-based algorithm, the accuracy of the detection reaches an improvement of 12.9% by the method proposed in [32] when SNR is 5 dB.

In order to further eliminate the problem that the amplitude of each spectral line after FFT is affected by noise, [33] presented a new endpoint detection method called Adaptive Band-partitioning Spectral Entropy (ABSE) algorithm combined with the Adaptive Band Selection (ABS) algorithm [34]. The idea of sub-band spectral entropy is to divide one frame into a number of sub-bands, then the spectral entropy of each sub-band is calculated.

The noisy speech signal in time domain is $x(n)$, $x_i(m)$ denotes the *i-th* frame of the speech signal after windowing and framing, its DFT is denoted as the following equation:

$$X_i(k) = \sum_{m=0}^{N-1} x_i(m)\exp(-j2\pi km/N) \tag{6}$$

where $X_i(k)$ denotes short-time Fourier transform of $x_i(m)$, and the energy of each component is:

$$Y_i(k) = |X_i(k)|^2 \tag{7}$$

The normalizing spectral probability density is given as follows:

$$p(k,i) = \frac{Y_i(k)}{\sum_{l=0}^{N/2} Y_i(l)} \quad k = 0, 1, \ldots\ldots, N/2 \tag{8}$$

And the information entropy of the first half of each frame can be calculated as:

$$H(i) = -\sum_{k=0}^{N/2} p(k,i)\log\, p(k,i) \tag{9}$$

where $H(i)$ denotes the *i-th* spectral entropy.

Dividing one frame into a number of sub-bands can avoid the interference of the noisy spectral magnitude. Assuming that each sub-band is composed of four lines, so there are N_b sub-bands in total, then the sub-band energy of the *m-th* line in the *i-th* frame can be denoted as Eq. (10).

$$E_b(m,i) = \sum_{k=(m-1)*4}^{(m-1)*4+3} Y_i(k) \quad 1 \le m \le N_b \tag{10}$$

Accordingly, the probability of the sub-band energy $p_b(m,i)$ is denoted as follows:

$$p_b(m,i) = \frac{E_b(m,i)}{\sum_{k=1}^{N_b} E_b(m,i)} \quad 1 \le m \le N_b \tag{11}$$

Then the sub-band spectral entropy $H_b(i)$ can be denoted as follows:

$$H_b(i) = -\sum_{m=1}^{N_b} p_b(m,i)\log\, p_b(m,i) \tag{12}$$

The features related to the spectral entropy in frequency domain for speech endpoint detection are described, especially the principle of sub-band spectral entropy. Next, we introduce another feature in frequency domain, that is spectral variance. It contains two aspects of information. One is the degree of ups and downs between each band in one frame. The other is the information of the energy of one frame. The higher the energy of the speech signal is, the more sharply it changes and the bigger the spectral variance is. On the contrary, the energy of the noise signal is low and changes slowly, which means the spectral variance of noise is small. So we can discriminate the speech segment from the noise signal based on the above characteristic. The mean value of the signal by FFT is described as Eq. (13).

$$E_\mathrm{i} = 1/N\sum_{\mathrm{k}=0}^{N-1} |X_i(k)| \tag{13}$$

The spectral variance is:

$$D_i = \frac{1}{N-1}\sum_{k=0}^{N-1} [|X_i(k)| - E_i]^2 \tag{14}$$

Spectral variance includes not only the fluctuation degree between each frequency band, but also the short-time energy characteristic of each frame signal. But for speech signals in real life, they always encounter the interference of the impulse noise, resulting in poor detection results. In order to improve the detection rate, a new method was proposed in [35], which employed a low-variance spectrum estimate and determined an optimal threshold based on the estimated noise statistics [36]. This method incorporated a low-variance spectrum estimation technique and a method for determining an adaptive threshold based on the noise statistics. These innovations in [36] resulted in a statistical test that is simple and elegant, whilst maintaining a high detection rate and a low error rate. Besides, considering the idea of introducing the sub-bands would make the algorithm robust to noise, a modified endpoint detection algorithm based on sub-band frequency variance was proposed in [37]. Spectrum was divided into several sub-bands, and the variance of each frame was computed. Finally, the consequence was smoothed through a median filter. The experimental results in [37] show that this modified algorithm can improve the performance of the endpoint detection in low SNR environments. In the calculation of the sub-band frequency variance, the length of each frame is N. There are totally $(N/2+1)$ spectral lines, after *FFT*, we can get the amplitude of the *i-th* frame spectral in Eq. (15).

$$X_i = \{X_i(1), X_i(2), \ldots\ldots, X_i(N/2+1)\} \tag{15}$$

Then we divide the spectrum X_i into several non-overlapping narrowband sub-bands.

$$XX_i(m) = \sum_{k=1+(m-1)p}^{1+(m-1)p+(p-1)} |X_i(k)| \tag{16}$$

where i is frame index, m is sub-band index, k is spectral line index and p is spectral line number of each sub-band.$XX_{\text{i}}(m)$ represents the magnitude of the *m-th* band of the *i-th* frame.

Suppose:

$$XX_i = \{XX_i(1), \ldots\ldots, XX_i(q)\} \tag{17}$$

where q is the number of sub-band.

Then the mean value of the sub-bands of the *i-th* frame is:

$$E_{i,1} = 1/q \sum_{k=1}^{q} XX_i(k) \tag{18}$$

The variance is:

$$D_{i,1} = \frac{1}{q-1} \sum_{k=0}^{q} \left[|XX_i(k)| - E_{i,1}\right]^2 \tag{19}$$

where $E_{i,1}$ denotes the sub-band mean of each frame, $D_{i,1}$ denotes the sub-band frequency variance of each frame.

3.2.2. Simulation implementation

Here we choose the same speech signal as in Section 3.1.2. Fig. 3 shows the results of sub-band frequency variance based method. As shown in Fig. 3, this algorithm can detect every word in the sentence, and can smooth the inter-word gaps when SNR is 5 dB. For the temporary pauses between the words such as 'jue' and 'ding', this algorithm still works. Due to a short pause between 'xiao' and 'lv', the two words are detected as a speech segment. However, the algorithm cannot accurately detect the starting and ending points of the adjacent words such 'ze' and 'de' because of the inter-word interferences caused by noise.

3.2.3. Sub-conclusion

With the decrease of SNR, the misjudgment rate of the adaptive sub-band spectral entropy method is also increased. To enhance the efficiency of speech endpoint detection, a positive constant K [38] was introduced into the calculation of basic Band-partitioning Spectral Entropy (BSE) to get an improved negative spectral entropy. Combined with the ABS method, a novel feature parameter called Adaptive Band-partition Constant Negative Spectral Entropy (ABCNSE) was achieved in [38]. Speech endpoints could be accurately detected through this feature. The experimental results in [38] reveal that this method is robust and valid under lower SNRs. In order to improve the detection accuracy as well as enhance the robustness of the endpoint detection algorithm in noise environments, in 2016, [39] improved the traditional spectral entropy algorithm and proposed two new endpoint detection parameters. Among them, the spectrum entropy based on critical band took both perceptual characteristics of human auditory system and distribution differences between speech and noise signals in frequency domain into account, as well as the minus frequency-domain energy parameter paid attention to the energy differences between speech frames and silence frames in frequency. Experimental results in [39] show that the endpoint detection algorithm has better discrimination for speech frames and silence frames and can carry out better accuracy than other conventional endpoint detection algorithms under low SNR environments, especially in the case of −5dB babble noise, the accuracy can be improved by more than 5% than the energy-entropy method in [39].

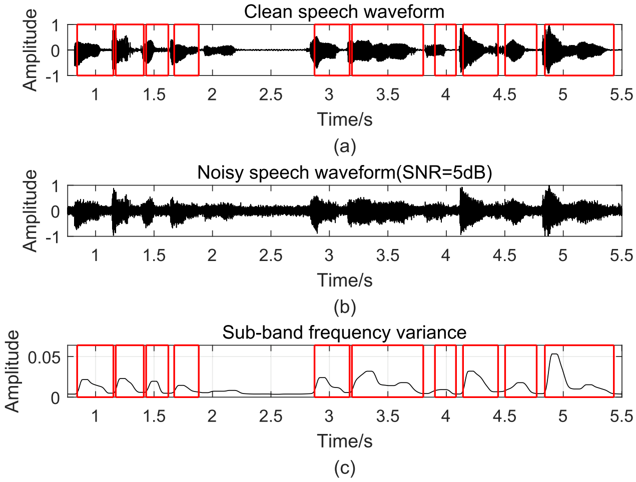


Fig. 3. (a) Waveform of speech "Fang fa de xuan ze, jue ding le ni de ban shi xiao lv"; (b) Waveform of noisy speech "Fang fa de xuan ze, jue ding le ni de ban shi xiao lv" at 5 dB; (c) Speech endpoint detection result using sub-band frequency variance method. The red blocks are representations of the detected speech segments. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Previous studies have proved that the endpoint detection using basic spectral variance becomes difficult and inaccurate when the SNR is very low. As a result, some improved algorithms based on the basic spectral variance are gradually emerging. A method was proposed in [40], which combined the adaptive sub-band spectral variance with spectral subtraction filter and integrated their advantages. The noisy speech was firstly enhanced using the spectral subtraction method. Then an adaptive sub-band selection spectral variance was used to detect speech endpoints. The experimental results in [40] show that it can accurately detect speech endpoints, reduce the computational complexity of the sub-band spectral variance and improve the system efficiency. Moreover, its robustness is better than the traditional spectral variance algorithm.

The above algorithms are mostly based on short-time characteristics of speech, but the long-term speech information is not fully considered. In order to make better use of the long-term characteristics of speech, scholars have proposed several methods based on long-term power spectrum variation of speech signals. For the sake of making up for the shortcoming that the linear long-term information does not fully utilize the acoustic characteristics of speech signal, the long-term information based on auditory filter bank was proposed in [41], where the correlation information between adjacent frames was fully utilized for endpoint detection. Though the detection performance in [41] was improved, the selection of the thresholds during the decision process directly could influence the detection results in different noise environments. In order to improve the detection performance at very low SNRs without selecting thresholds, in 2018, six types of long-term information based on auditory filter banks were extracted in [42] through the nonlinear spectral decomposition with three different auditory filters. Further, an adaptive speech detection algorithm based on these types of long-term information was proposed in [42]. Without additional training data, this algorithm used the data selected from the test signals according to long-term information to train a speech/non-speech classifier, and classified the current test signals using the speech/non-speech classifier frame by frame. Experimental results show that the algorithm can perform well even in a low SNR of -10 dB.

3.3. Features in time-frequency combination

The previous studies found that time-domain endpoint detection algorithms are simple and can be implemented easily, but have poor performance under the background of noise environments. The frequency-domain endpoint detection algorithms are robust at the expense of computational complexity. Considering the endpoint detection methods in time domain and frequency domain both have their own advantages, scholars began to explore new endpoint detection methods using the combined features in time domain and frequency domain.

3.3.1. Concepts

In order to precisely locate the endpoints of the input speech signal excluding non-speech segments, a novel approach combining the time-domain features and the spectral entropy feature for speech endpoint detection was proposed in [43]. In this proposed method, time-frequency enhancement and spectral entropy feature were used together. Firstly, noisy speech was enhanced using the spectral subtraction method to remove the additive noise in frequency domain. Then in time domain, a weight function built by short-time energy and ZCR was used to remove noise produced by the spectral subtraction. Finally, the spectral entropy-based method was used to detect the endpoints. The proposed algorithm in [43] is shown to be well suited for the endpoint detection and is very robust to different types of noise, especially for low SNRs. Furthermore, the algorithm has a low complexity and is suitable for the real-time digital signal processing system. Based on the study of [43,44] proposed an improved endpoint detection algorithm, which took a real-time multistage detection on the speech from telephone channel using the frequency spectrum entropy and the Short-time Energy-zero Value (the product of short-time energy and ZCR) as the decision-making parameters. The algorithm in [44] was proposed adequately to solve the issue that the dual-threshold and the spectral entropy cannot judge accurately in real time.

To solve the problem that traditional algorithms are in low detection accuracy, random oscillation and are easy to misjudge under the non-stationary noise, in 2013, [45] proposed a new algorithm combining the energy difference, periodicity, ZCR, and the spectral differences between different frames. Results show that the detection accuracy of this method is higher than that of the traditional algorithms, especially in the case of background noise environments.

In addition to the methods mentioned above, a more effective sub-band detection algorithm based on time-frequency features for mandarin was proposed in [46]. The proposed algorithm includes two parts: the crosswise detection and the lengthwise detection. Besides, the energy detection and pitch detection were also used in [46]. For better performance, the dual-threshold criterion was used to reduce the misjudgment rate of the detection. Along with the whole process, the misjudgment correction and noise updating were done. The experimental results in [46] indicate that the proposed algorithm can detect the voice segments effectively in non-stationary and low SNR noisy environments. But for more complex noise environments such as the burst noise, there is still room for improvements. Considering the sub-band energy [47,48] can effectively suppress the interference of the burst noise. Stated thus, a new speech feature called sub-band energy entropy ratio was proposed in [49], which can be denoted as Eq. (20).

$$SHE(i) = \sqrt{1 + |SE(i)/H_b(i)|} \tag{20}$$

where $SHE(i)$ denotes the sub-band energy entropy ratio of the *i-th* frame of signal, $SE(i)$ represents the sub-band energy of each frequency band in each speech frame, and $H_b(i)$ represents the sub-band spectral entropy. The experiment in [49] shows that the accuracy of the proposed algorithm is better than that of the conventional sub-band spectral entropy algorithm under the four different SNRs including 0, 5, 10 and 15 dB. Especially, in white noise, the average accuracy rate increases by 13.85%; in Babble noise, it increases by 17.54%. The simple calculation can effectively lighten the calculation load of the system and reduce the processing time. Further, focusing on the use of combined features for endpoint detection, [50] constructed a compound parameter D using the spectral entropy, short-time energy, ZCR, and the linear prediction error. Then $D/D_{\max}$ was used to determine the speech, non-speech, and silent frames. The results in [50] show that this method can accurately detect speech signal endpoints while testing with TIMIT database, however, the threshold values are determined empirically.

3.3.2. Simulation implementation

Here we choose again the same speech signal as in Sections 3.1.2 and 3.2.2 for experiment. The results of the sub-band energy entropy ratio algorithm in time-frequency domain are shown in Fig. 4. From Fig. 4, we know that this algorithm performs poorly in the detection of the words whose initials are unaspirated sounds such as ‘b’, ‘d’, ‘z’, ‘j’ at 5 dB. Meanwhile, the detection performance will be degraded if the initials are aspirated sounds such as “x” or low energy syllables such as ‘f’, ‘b’, ‘s’. Especially when the starting points are low-energy initials, this algorithm tends to misjudge these initials as the noise or the finals of the previous words and thus will reduce the detection accuracy.

3.3.3. Sub-conclusion

At present, the endpoint detection methods based on the features in combination of time domain with frequency domain are almost based on the above-mentioned original features to construct composite parameters for threshold decision. In low SNR and uncomplicated noise environments, the efficiencies of the above endpoint detections are significantly higher than those of the traditional algorithms. However, the detection accuracy often decreases sharply in more complex and changeable noise environments. This type of combined feature in the process of decision is generally judged by the threshold method, and the selection of the thresholds obtained empirically has a great influence on the detection accuracy.

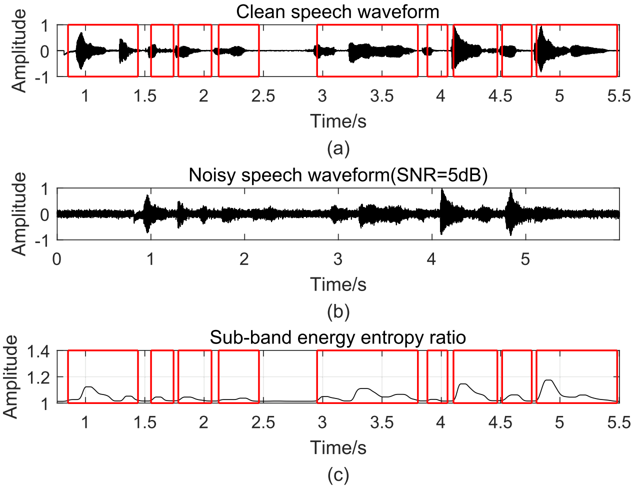


Fig. 4. (a) Waveform of speech "Fang fa de xuan ze, jue ding le ni de ban shi xiao lv"; (b) Waveform of noisy speech "Fang fa de xuan ze, jue ding le ni de ban shi xiao lv" at 5 dB; (c) Speech endpoint detection result using sub-band energy entropy ratio algorithm. The red blocks are representations of the detected speech segments. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.4. Features in cepstrum domain

Based on the development of time-frequency spectrum analysis and psychoacoustics technology, the features in cepstrum domain began to appear. Conceptually, the cepstrum can be seen as a rate of the changes in different spectrum bands of signal. Some studies show that the features in cepstrum domain can be used for various speech endpoint systems.

3.4.1. Concepts

The selection of the features in cepstrum domain plays a very important role on speech endpoint detection. The commonly used features are the LPCC and MFCC. The system function of the vocal tract model obtained by linear prediction analysis is:

$$H(z) = 1/(1 - \sum_{i=1}^{p} a_i z^{-i}) \tag{21}$$

The impulse response is $h(n)$, on the basis of the homomorphic processing, suppose $\hat{h}(n)$ is the cepstrum of $h(n)$, we can know:

$$\hat{H}(z) = \log H(z) \tag{22}$$

$$\hat{H}(z) = \sum_{n=1}^{+\infty} \hat{h}(n) z^{-n} \tag{23}$$

The above formulae can be used to calculate $\hat{h}(n)$ directly from the prediction coefficients a_i. LPCC is easy to realize, but also can be easily influenced by additive noise and its description ability of the consonants is poor. Although LPCC and its related parameters have achieved great success in modeling vowels, they are not suitable for modeling the nasals and fricatives, and the reliability of LPCC depends on the noise environments.

The concept of MFCC was proposed in Davies et al. [12], where MFCC was calculated by converting the spectrum into a nonlinear spectrum based on the Mel frequency standard, and then converting it into the spectrum domain. MFCCs are the amplitudes of the Discrete Cosine Transformation (DCT) spectrums. Before long, the cepstrum distance was pointed out in [51] to be used as a decision parameter for speech endpoint detection. The cepstrum parameter is obtained by the homomorphism analysis of the speech signal, and the mean square distance of the cepstrum can reflect the differences between the two signal spectrums. The cepstrum distance detection method judges the endpoints via the trajectory of the cepstrum distance between each signal frame and noise frame.

3.4.2. Simulation implementation

Again, here we choose the same speech signal, whose context is "Fang fa de xuan ze, jue ding le ni de ban shi xiao lv"(In Chinese Pinyin). Fig. 6 shows the results of the classic MFCC cepstrum distance algorithm. By analyzing Fig. 5, we can find that the dynamic variability range of the short-time MFCC cepstrum distance is so large that it puts forward a higher requirement for the threshold setting at the decision-making stage when SNR is 5 dB. This algorithm has a poor detection effect for the syllables starting with low-energy initials, and it cannot eliminate the influence of the inter-word gaps on the endpoint detection. For instance, the latter 'de' has not been detected in Fig. 5.

When the SNR is not ideal, the serious spectral distortion of the speech signal itself will bring difficulties to the threshold estimation. Besides, it is difficult to distinguish between speech signals and noise signals whose cepstrum distances are similar to some speech signals like the non-stationary noise such as the sound of opening the door et al.

3.4.3. Sub-conclusion

Subsequently, many scholars have come up with many improved algorithms. A novel statistical endpoint detection algorithm based on the cepstrum coefficients and their moments was proposed in [52]. In this method, the moment ratios of the speech segments and the silent segments were used to evaluate a threshold measure for differentiating between silent and speech segments of the conversation. To make it robust in noise environments, it gradually tuned the threshold to adapt to the dynamic background noise. The simulation results in [52] show the compression ability of this method in various environments.

In addition to the features mentioned above, in order to solve the problem that the accuracy of speech endpoint detection using the traditional method dramatically declines, some scholars have made some improvements on MFCC features. On the basis of MFCC, it was mentioned that the endpoint detection algorithm using

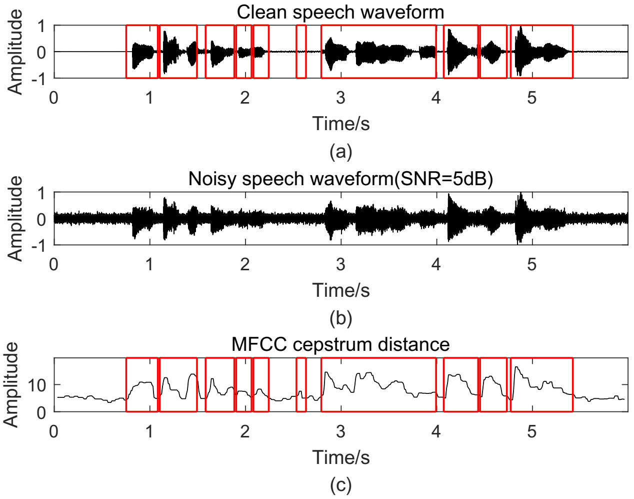


Fig. 5. (a) Waveform of speech "Fang fa de xuan ze, jue ding le ni de ban shi xiao lv"; (b) Waveform of noisy speech "Fang fa de xuan ze, jue ding le ni de ban shi xiao lv" at 5 dB; (c) Speech endpoint detection result using MFCC cepstrum distance. The red blocks are representations of the detected speech segments. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

MFCC similarity [53] could achieve a better detection effect at low SNR circumstances compared with the traditional short-time energy method. Besides, a new algorithm based on MFCC distance with an order statistics filter was proposed in [54]. First, the MFCC for each frame of the signals was extracted. Then, the background noise was estimated using the first sixteen frames. Finally, the MFCC cepstrum distances between each frame and the background noise were calculated. An order statistics filter was applied to a sequence of the estimated cepstrum distances to obtain the weighted cepstrum distance of each frame. The detection was based on the weighted cepstrum distance. The experimental analysis carried out on the TIMIT speech corpus in [54] shows that the proposed algorithm performs well with white noise, pink noise, car noise, and fighter noise even in low SNR conditions. However, in the case of very low SNRs for specific noise, the traditional cepstrum distance method is often difficult to accurately judge the starting and ending points of the speech segments, so a new method based on the multi-window spectrum estimation spectral subtraction and the improved cepstrum distance was proposed in [55], which could be useful to detect the endpoints of the speech in pink noise when SNR is −8 dB. However, there is no in-depth research for other complex environments.

3.5. Features in other transform domains

Apart from the features in time domain, frequency domain and cepstrum domain, scholars have proposed some features in other transform domains for the non-stationary speech signals such as wavelet transform, Empirical Mode Decomposition (EMD) and Teager energy.

As described in [56], the speech segments and the noise segments can be distinguished based on the theory that in every region of the wavelet transform, the noise energy changes slowly. The experiment in [56] shows that this means could distinguish the speech and noise in both high SNR and low SNR situations. Subsequently, an algorithm combining the wavelet transform with Support Vector Machine (SVM) was proposed in [57]. The wavelet transform can refine the signal (function) in multiple aspects gradually through local analysis, stretching and panning on space (time) frequency. Ultimately, it realizes the time subdivision at high frequency points and the frequency subdivision at low frequency points, respectively. SVM defines a hyperplane through a known kernel function so as to divide the given points into two predefined classes. The computing speed of the SVM algorithm, however, is low. Thus, a multiple kernel SVM method for the multiple features was proposed in [58] which aimed to overcome this shortcoming. This method adapted the multiple kernel learning thought to an efficient cutting-plane structural SVM solver. The experimental results in [58] shows that the proposed method not only can lead to better global performance by taking the advantages of the multiple features but also has a low computational complexity.

We can use instantaneous frequencies as the basic parameter and use intrinsic mode components as the basic signals of the time domain to construct a new time-frequency analysis system. In this way, any signal can be decomposed into some intrinsic mode functions. This method is called EMD. EMD was a method proposed by Huang et al. in 1998 to analyze and process continuous nonlinear non-stationary signals [59]. A new speech endpoint detection method based on EMD was proposed in [60]. Firstly, the IMF1 and IMF2 were discarded after EMD, and then some IMFs components that could be used in the endpoint detection were selected to reconstruct the speech signal by comparing their means and variances. The experiment in [60] shows that the proposed method remains high detection rate of 84.2% compared with the dual-threshold method. Though there is noise existing in the speech signal, the proposed method in [60] could still implement detecting with detection rate being 85.9% while the traditional dual-threshold method failed.

The Teager energy of a discrete-time system is defined as follows:

$$T[x_i(m)] = [x_i(m)]^2 - x_i(m+1)x_i(m-1) \quad m = 1, 2, \ldots, N \tag{24}$$

In 2016, a power spectrum bias algorithm [61] based on Teager energy was proposed to enhance the discernibility between the speech and noise signals, thus eliminating the problem of the traditional power spectrum deviation. Meanwhile, this algorithm took the probability of absent speech derived on the basis of the likelihood ratio of the Teager energy as a smoothing parameter to further revise the power spectrum deviation in various noise environments.

Based on the above features in different transform domains, many improved algorithms have emerged. In 2014, an improved methodology based on the ensemble EMD algorithm and Teager kurtosis to avoid the defect of the EMD in mode mixing was proposed in [62]. The Teager energy operator was used to track the modulation energy of each IMF, decomposed by ensemble EMD. The root power function and the order statistics filter were used on Teager kurtosis for feature extraction. This method could be implemented over suitable thresholds which could be automatically estimated by tracking the minimum of the extracted feature values. The experiment in [62] shows that the proposed algorithm can achieve comparable results at high SNRs. For low SNR conditions, it is able to maintain lower error detection ratio and higher detection ratio, compared with those of the original algorithms. Furthermore, according to different application requirements, time-frequency wavelet transform and wavelet entropy energy ratio are also used for speech endpoint detection. Compared with the traditional methods, the principle of wavelet energy entropy is easy to implement. Additionally, it can be applied to random signal processing. Note that the wavelet energy entropy fails to detect the speech endpoints in a noise environment.

4. Decision method

Generally speaking, the endpoint decision algorithms can be roughly divided into two categories: one is the threshold-based decision method. This kind of method extracts the eigenvalues of each speech frame to be measured first, then compares the eigenvalues with the preset thresholds to execute decision. It is so simple and fast that it can meet the requirements of real-time system. The other one is the pattern matching decision algorithm. Although this sort of method is complex to be realized for acquiring more training data, it has higher accuracies than the first kind of methods averagely.

4.1. Threshold decision method

The selection of speech features directly determines the accuracies of the threshold decisions. Threshold decision methods are mainly applied to pure speeches. In view of the above features in various domains, a dual-threshold decision algorithm of both dual-parameter and single-parameter has been proposed. In dual-threshold decision algorithm, a higher threshold of a certain feature is selected firstly to perform a rough decision. That is, the parts higher than this threshold are set to the speech segments. Then a lower threshold is selected, and the points that intersect the lower threshold are searched from the first judged junction points to both sides of them. If the decision threshold is set properly, the algorithm can perform better detection of the speech signals. However, the threshold values need to be determined empirically. The threshold has a great impact on the performance of the whole speech endpoint detection. Therefore, how to set a reliable and accurate threshold that can be adaptive to the background noise environments is a problem to be solved in this algorithm. Afterwards, some scholars have proposed an adaptive algorithm for dynamically adjusting the threshold, which can adapt to the detection of speech endpoints under different SNRs. However, the experimental results show that the threshold decision algorithm has poor detection performance in low SNR and other non-stationary noise environments. Moreover, the short-time impulse noise whose energies exceed the threshold value, such as the people's smacking sound, a popping sound or a breathing sound, are not be taken into account, which needs to be considered in the future research.

4.2. Pattern matching method

Pattern matching method works well in a noise environment. However, this algorithm is unsuitable for the real-time implementation because of its high computational complexity and the requirement for a large amount of prior knowledge. This sort of method is generally classified into the Gaussian mixture model method and the high-order statistics method.

The principle of the decision based on the Gaussian mixture model is to classify the extracted feature vectors into several classes, and to assume that both inter-classes and intra-class vectors are independent and can obey the same normal distribution, then the general distributions of the speech and noise feature vectors are obtained by adding the normal distribution of multiple classes according to a certain weight value. Afterwards, the speech and the noise models are respectively established according to the training mean value, covariance and other parameters. According to the principle of the maximum posterior probability, each frame of the input signal can be determined as a speech or a noise frame, while the model parameters can be updated.

For the speech signals affected by strong background noise, some methods based on the statistics [63,64,65] have been proposed. A new detection algorithm that was based on the statistical models and the empirical rule-based energy was presented in [63]. Specifically, it needs two steps to separate the speech segments from the background noise. For the first step, the possible speech endpoints were detected efficiently using the empirical rule-based energy detection algorithm, for the second step, the endpoints were aligned to their optimal positions using a new Gaussian mixture model-based multiple-observation log likelihood ratio algorithm. The results in [63] show that this algorithm can achieve better performance in various noise scenarios. A decision method based on the statistical likelihood ratio adaptive threshold was considered in [65], which introduced a kind of method based on the adaptive threshold of statistical likelihood ratio, and compared the likelihood ratio of the current frame with the threshold to perform the decision of speech frames and non-speech frames. Besides, in 2016, an improved method which combined the statistical noise suppression method with the convolutional neural network was proposed in [66]. Noise signal is always assumed to satisfy and characterized by Gaussian distribution in most of the statistical model based VAD algorithms. However in low Signal-to-Noise Ratio (SNR) conditions, the assumption of noise does not always hold in practice. For further improving the robustness of VAD, an enhanced speech based method is proposed [67] where Laplacian distribution is used to model the remained noise and Gaussian mixture model is used to characterize the Discrete Fourier transform (DFT) coefficients of reconstructed speech in enhanced speech. The above statistical methods are complex, so it is still challenging while obtaining the robustness of the algorithm.

Apart from the above statistical methods, there are other decision methods by training the Hidden Markov Model, SVM and neural network. Here we mainly lay emphasis on the algorithms based on neural network. In fact, the threshold decision method is widely used during the decision stage. However, due to the uncertainty of threshold values, the detection effect is not ideal in complex environments with low SNRs. To solve this problem, algorithms based on neural network have emerged. The purpose of speech endpoint detection based on neural network is to use the extracted features from speech signals as the input of the trained neural network, and then to identify, analyze and judge the speech signal through the neural network so as to find the starting and ending points of the speech segments. Back in the early 1990s, Crippa et al. from the University of Pittsburgh creatively applied the artificial neural network algorithm to speech endpoint detection. Thereafter, neural network has always been attracting researchers who would like to realize the target of the strong anti-interference ability, strong eigenvalue generalization ability and high robustness under lower SNRs or actual presence of the non-stationary noise environments. In 2012, it was found that using the large-span feature could significantly improve the robustness in real-life noise environments in [68]. From the above exposition in [68], a data-driven method based on Long Short-Term Memory-Recurrent Neural Networks (LSTM-RNN) was presented in [69], which was evaluated on unseen synthetically mixed test data as well as a real-life test set consisting of four full-length Hollywood movies. The experimental results in [69] show that a frame-wise Equal Error Rate (EER) of 33.2% is obtained for the four movies and an EER of 9.6% is obtained for the synthetic test data at a peak SNR of 0 dB. However, when the characteristics of these speech signals are similar to the characteristics of noise, such as the music and ambient noise in low SNRs, its performance will degrade. In order to solve this problem, an algorithm based on the LSTM of the modulation spectrum was proposed in [70] aiming to solve the issue that the fricative sound could be misjudged as noise. The detection accuracy of this algorithm in [70] increases by 3.5% under the condition of different noise and different SNRs. Due to the fact that VAD usually works with a dynamic decision threshold and ROC curve is a global evaluation metric of VAD, so a method of optimizing the area under ROC curve (AUC) by DNN [71] is proposed for VAD. Results show compared with the common method of optimizing the minimum squared error by DNN, the proposed method can result in higher performance.

At present, most of the mainstream algorithms combine various forms of neural network algorithms with other feature transformations to perform the decision process. For example, a new algorithm based on boosted Deep Neural Networks (bDNNs) was described in [72]. The proposed algorithm first generated multiple base predictions for a single frame from only one DNN and then aggregated the base predictions for a better prediction of the frame. Moreover, a new acoustic feature, Multi-Resolution Cochleagram (MRCG) [73] was used to concatenate the cochleagram features at multiple spectrotemporal resolutions. The experimental results in [73] show that the MRCG feature performs the best among the 16 evaluated features including the time domain feature such as the autocorrelation, the cepstrum domain feature such as the MFCC and so on under the condition of six different kind of noise at −5 dB. Based on the results of [73], a new method called Multi-Resolution Stacking (MRS) was presented in [74], which used a new base classifier-bDNN and a newly introduced acoustic feature-MRCG. Here this method explored contextual information by machine learning methods at three levels. At the top level, an ensemble learning framework named MRS was employed, which was a stack of ensemble classifiers. At the middle level, a base classifier in MRS named bDNN was described. At the bottom level, the MRCG feature was employed, which incorporated the contextual information by concatenating the cochleagram features at multiple spectrotemporal resolutions. The experimental results in [74] show that the MRS-based method outperforms other methods by a considerable margin. Moreover, when trained on a large amount of noise types and a wide range of SNRs, the MRS-based method demonstrates surprisingly good generalization performance on the unseen test scenarios, approaching the performance with noise-dependent training. In addition, [75] improved the use of contextual information by using an adaptive context attention model (ACAM) with a novel training strategy for effective attention. This strategy weights the most crucial parts of the context for proper classification.

For the moment, almost all the speech endpoint detection algorithms only consider the amplitude information of the signal itself but the phase information that accounts for half of the signal. In 2017, an endpoint detection algorithm based on deep neural network analyzing comprehensively the signal amplitude phase information was showed in [76]. The experimental results in [76] show that the error rate is greatly reduced compared with the DNN algorithm based on the signal amplitude. For better evaluating the performance of deep learning based VAD, researchers in [77] introduced a new dataset and evaluated GMM and DNN on it. For GMM, two separate mixtures were used to model speech and nonspeech. For neural networks, a softmax layer was used at the end of the network, with two neurons which represent speech and non-speech. The network was trained using stochastic gradient descent to minimize cross-entropy loss. Finally, an accuracy of 81.61% was yielded for GMMs, and 85.18% for DNNs. However, the endpoint detection algorithms based on the neural network generally have the problems of slow training speed and high complexity. How to reduce the computational complexity while ensuring the detection accuracy is the direction of future researches.

5. Simulation experiment

Among the above-mentioned detection algorithms, we select several classical algorithms to carry out the simulation experiments and compare the endpoint detection accuracies.

5.1. Preprocessing

5.1.1. Framing

Speech is stationary in short time, so framing is the first performed. In this process, there are often overlapping parts between two adjacent frames which can smooth the characteristic parameters. Frame shift is defined as the displacement of the two adjacent frames.

The discrete signal is:

$$x = (x(1), x(2), \ldots, x(k)) \tag{25}$$

Generally, frame size is about 10–30 ms. The K data is divided into *L* frames, that is there are *K/L* sampling points in each frame. Frame shift is set to 1/2 or 1/3 of the frame size.

5.1.2. Windowing

Framing is equivalent to multiplying a finite length window function. Suppose $x(n)$ is the sampled speech signal, then we can get the windowed speech signal as:

$$y(n) = \sum_{n=-\infty}^{\infty} x(m)\omega(n-m) \tag{26}$$

Window functions generally have low pass characteristics. Studies show that the bandwidths and the spectrum leakages of different window functions are different. The mutation of the rectangle window in time domain leads to serious spectrum tailing and slow convergence. Therefore, the leakage of the spectrum can be reduced by improving the shape of the window function. Hamming window is a commonly used window function:

$$\omega(n) = 0.54 - 0.46\cos(2\pi n/N - 1), 0 \le n \le N \tag{27}$$

where *N* is the total number of sampling points, when n is equal to other values except for the values ranging from 0 to *N*, the window function is 0.

5.2. Experimental preparations

The experimental data in FLAC format is selected from the European Broadcasting Union (EBU) professional lossless audio database. The data is recorded on two 16 kHz channels and stored as 16 bits PCM with a sampling rate of 44.1 kHz. FLAC is a compression method designed specifically for the features of audio and it does not destroy any of the original audio information. For ease of the experimental test, we select the audio signals which can last for 3–5 s, including the sine wave signals of different frequencies, the audio from various musical instruments (such as the trumpet, harp, castanets, xylophone, piano, guitar, etc.) and the recordings from different gender testers. In the experiment, these audio signals are down-sampled to 16 kHz, the mono is selected, the frame size is 32 ms (for *FFT*), that is 512 samples, and a frame shift of 256 samples. In order to estimate the noise, the duration of the leading no-voice segment is set to 0.25 s according to the speech signal. Before the experiment, these test signals are first manually labeled ('0' for noise segments and '1' for speech segments). Then the experiments of the performance of different algorithms are conducted under four noisy environments, which are WGN, PINK, BABBLE and MACHINEGUN environments. Each noisy environment contains six SNR conditions (-5, 0, 5, 10, 15, 20 dB) respectively. Here we regard the detection accuracy as the evaluation parameter. The detection accuracy is defined as:

$$Accuracy = \frac{N_{num} - N_{err}}{N_{num}} \times 100\% \tag{28}$$

where N_{num} denotes the number of total frames, N_{err} indicates the number of error frames, which includes the number of the speech frames that are erroneously judged as noise frames and the number of the noise frames that are misjudged as speech frames.

5.3. Experimental steps

The specific procedure of the detection is presented in Fig. 6.

As shown in Fig. 6, all the essential steps involved in this detection are given below:

1) Remove the direct current component and normalize the amplitude;
2) Preprocess and set the initial parameters (if necessary, add noise to the signal);
3) Extract different features according to different algorithms;
4) Carry out the smoothing process, calculate the thresholds, and execute the dual-threshold decision for threshold decision methods, while training the neural networks for the neural network methods;
5) Compare with the manually calibrated signal and calculate the detection accuracy.

5.4. Algorithm demonstration

In order to better demonstrate and compare the traditional detection effects of different algorithms, we design a Graphical User Interface (GUI) implementation platform in the MATLAB2016. The Demo platform is showed in Fig. 7. This platform comprises

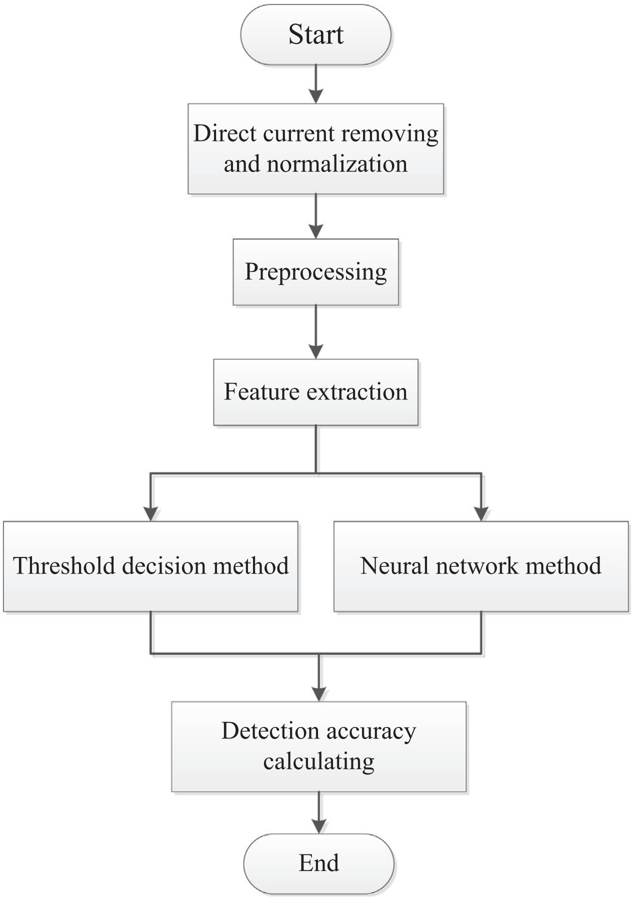

Start
Direct current removing and normalization
Preprocessing
Feature extraction
Threshold decision method
Neural network method
Detection accuracy calculating
End

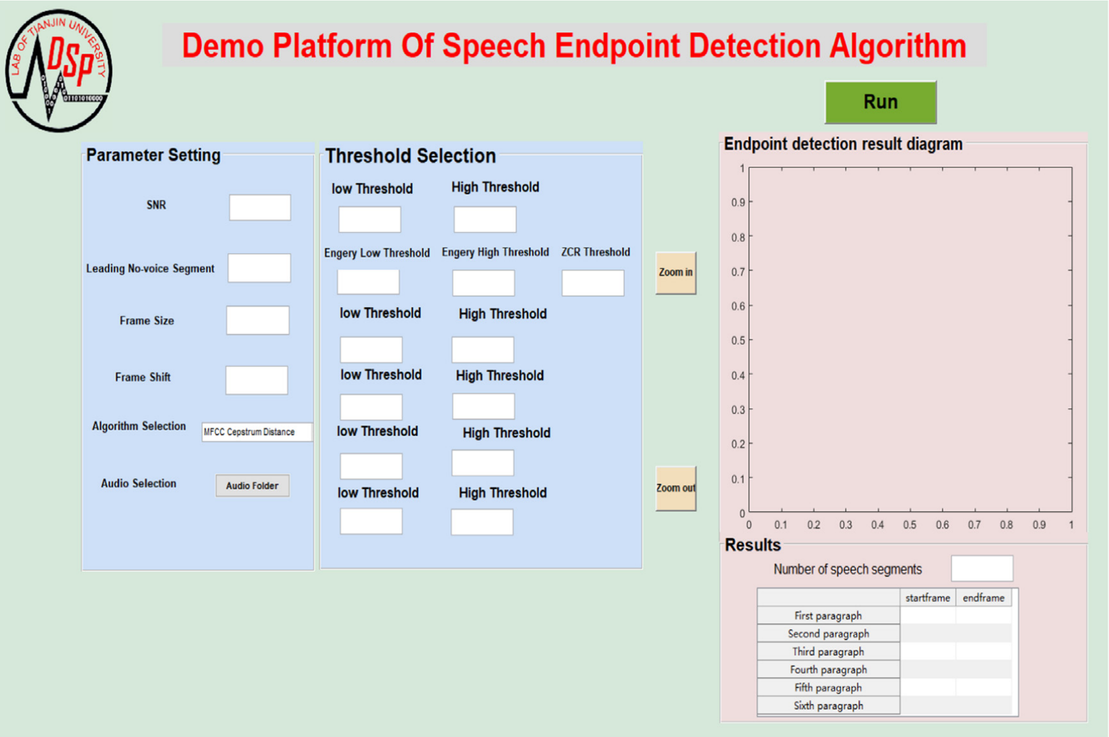
Fig. 6. The experiment flowchart.

three modules including the parameter setting, threshold selection and the result display. As shown in Fig. 7, users can set some initial parameters including the observed detection algorithm and the testing audio in parameter setting module. For different algorithm, the optimal threshold values can be set by comparing the detection results in the right diagram with different threshold values. The detection results will be displayed in the form of tables and graphical interfaces which can be zoomed in and zoomed out. GUI has the advantage of containing multiple algorithms together for an intuitive comparison and making a more convenient experience for users.

For example, the demo of the detection results is showed in Fig. 8 when the algorithm of MFCC cepstrum distance is chosen. In the simulation experiments, six algorithms based on various features are implemented on the GUI platform. These features include the short-time energy and the ZCR, autocorrelation, energy-zero ratio, spectral variance, sub-band energy entropy ratio and the MFCC cepstrum distance. Apart from the methods based on the above short time features, VAD methods based on the three mentioned long-term features LTSV, LSFM and LSVM are also included. Since neural networks are popular and important in current signal processing, here we take three mainstream neural network based VAD methods including BP, SVM and DNN for a more comprehensive comparison.

5.5. Results

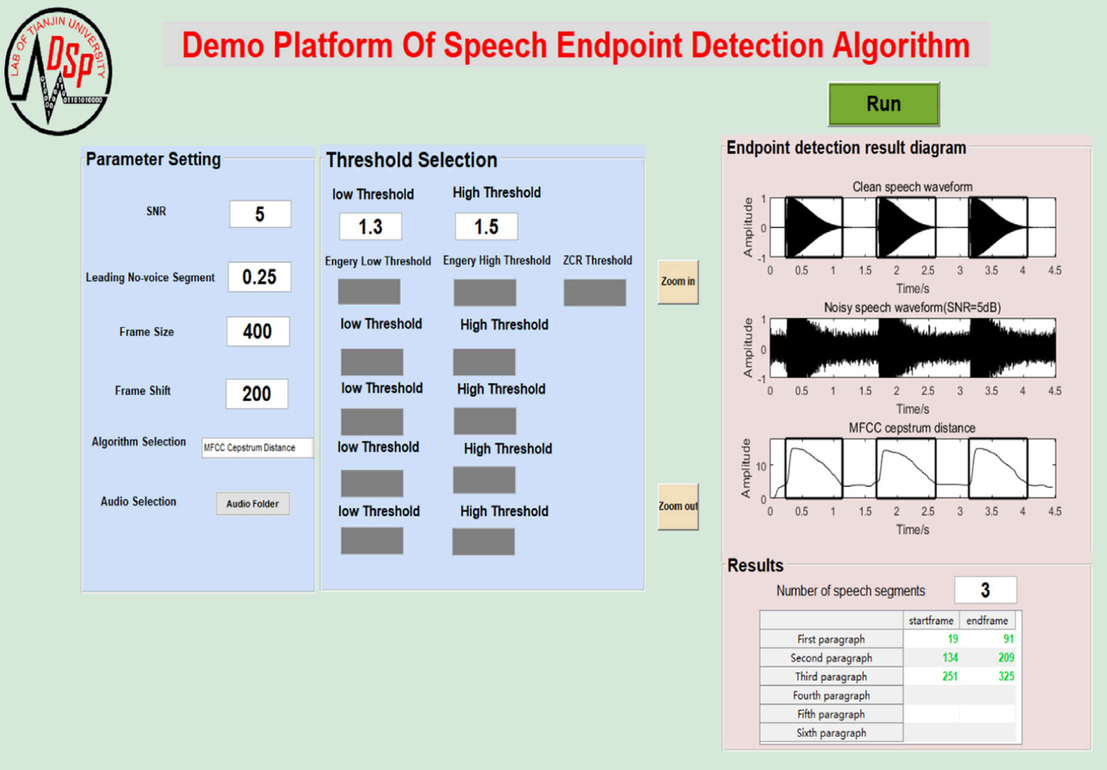
As we can see in Table 1 and Table 2, different accuracies have been achieved in the four different types of noisy environments, which are WGN, PINK, BABBLE and MACHINEGUN, with different SNRs. The obtained results are averaged in Fig. 9, *Algm1* to *Algm12* denote the algorithms of short-time energy with ZCR, autocorrelation, energy-zero ratio, spectral variance, sub-band energy entropy ratio, MFCC cepstrum distance, LTSV, LSFM, LSVM, BP network,



Algorithm Selection: MFCC Cepstrum Distance

	startframe	endframe
First paragraph		
Second paragraph		
Third paragraph		
Fourth paragraph		
Fifth paragraph		
Sixth paragraph		

Fig. 7. Demo platform of speech endpoint detection algorithm.



Demo Platform Of Speech Endpoint Detection Algorithm

Parameter Setting

SNR: 5

Leading No-voice Segment: 0.25

Frame Size: 400

Frame Shift: 200

Algorithm Selection: MFCC Cepstrum Distance

Audio Selection: Audio Folder

Threshold Selection

low Threshold: 1.3

High Threshold: 1.5

Results

Number of speech segments: 3

	startframe	endframe
First paragraph	19	91
Second paragraph	134	209
Third paragraph	251	325
Fourth paragraph		
Fifth paragraph		
Sixth paragraph		

Fig. 8. The demo of the MFCC cepstrum distance algorithm results.

Table 1
The endpoint detection accuracies of algorithms based on short-time features.

Noise Type	SNR	Short-time energy + ZCR	Autocorr-elation	Energy-zero ratio	Spectral variance	Sub-band energy entropy ratio	MFCC cepstrum distance
WGN	−5dB	76.80	85.29	84.79	90.77	91.52	77.83
	0 dB	77.81	87.78	89.03	92.27	94.76	85.14
	5 dB	78.30	89.03	89.08	94.26	95.76	88.25
	10 dB	81.30	92.27	90.27	96.01	97.51	90.22
	15 dB	82.12	92.52	90.77	96.51	97.76	93.71
	20 dB	82.45	92.77	92.77	97.26	97.90	94.79
AVE		73.61	89.94	89.45	94.51	95.86	88.32
PINK	−5dB	68.95	73.61	72.82	78.78	76.26	77.26
	0 dB	67.49	77.73	82.72	81.43	80.76	80.37
	5 dB	71.65	79.94	78.33	85.13	79.88	83.19
	10 dB	74.95	84.72	82	86.64	84.23	85.88
	15 dB	79.54	87.15	84.78	88.11	86.32	89.52
	20 dB	86.02	89.89	88.27	89.66	91.49	89.79
AVE		74.76	82.18	81.49	84.95	83.16	84.34
BABBLE	−5dB	66.45	75.66	71.27	76.45	77.23	74.91
	0 dB	74.97	78.21	73.84	81.17	80.01	78.30
	5 dB	78.01	82.49	76.23	82.56	83.16	79.70
	10 dB	82.24	85.92	77.08	82.85	84.21	83.34
	15 dB	83.31	85.84	81.85	84.91	86.78	85.03
	20 dB	85.75	88.39	85.43	88.45	90.61	86.98
AVE		78.45	82.75	77.62	82.73	83.67	81.38
MACHINEGUN	−5dB	69.92	72.52	74.32	73.20	76.33	82.44
	0 dB	74.65	74.83	73.71	77.72	80.57	84.00
	5 dB	78.79	80.09	74.38	79.36	81.75	83.67
	10 dB	79.61	84.20	76.87	82.18	83.80	84.84
	15 dB	85.24	85.82	80.41	82.78	86.36	86.98
	20 dB	86.19	87.81	82.72	85.61	87.61	86.99
AVE		79.07	80.88	77.07	80.14	82.73	84.82

SVM and DNN. Among these four types of noise, clearly we can conclude that the algorithms each can obtain a best average accuracy in WGN environments. While in the other three noise types, the accuracies are almost in a descending order of PINK, BABBLE and MACHINEGUN. Especially, the circumstance in each kind of noisy environment is almost the same as the other three noisy types. So here we pick just the typical noise type as the example to illustrate our results detailedly. Actually, the algorithms used in this paper can be generally divided into three categories. One is about the traditional short-time methods, one is about the

Table 2
The endpoint detection accuracies of algorithms based on long-term features and neural networks.

Noise Type	SNR	LTSV	LSFM	LSVM	BP	SVM	DNN
WGN	−5dB	90.66	88.75	91.57	80.87	78.55	76.98
	0 dB	95.27	95.21	96.73	84.45	84.25	82.62
	5 dB	97.66	97.98	98.56	86.35	85.84	87.11
	10 dB	98.37	98.72	98.76	88.74	88.57	90.63
	15 dB	97.84	97.96	96.44	89.63	89.94	92.37
	20 dB	97.46	97.61	96.44	90.14	90.16	92.84
AVE		96.21	96.04	**96.42**	86.70	86.22	87.09
PINK	−5dB	88.41	87.23	91.31	72.26	78.39	75.79
	0 dB	96.55	95.91	96.98	82.67	80.39	79.06
	5 dB	98.24	98.40	98.72	82.95	84.37	83.68
	10 dB	98.40	98.77	98.78	87.91	88.18	88.98
	15 dB	98.76	98.76	96.09	89.63	89.31	91.18
	20 dB	98.76	98.76	96.09	90.64	89.71	93.18
AVE		**96.52**	96.31	96.33	84.34	85.06	85.31
BABBLE	−5dB	78.54	78.14	78.68	76.98	78.28	75.79
	0 dB	81.67	81.86	85.58	79.82	80.56	77.91
	5 dB	86.19	86.69	90.33	82.04	81.64	83.98
	10 dB	90.27	91.35	92.38	86.24	85.16	87.62
	15 dB	82.89	92.53	93.25	87.66	87.04	90.59
	20 dB	82.89	93.01	93.25	89.71	88.69	91.90
AVE		83.74	87.26	**88.91**	83.74	83.56	84.63
MACHINEGUN	−5dB	76.09	76.96	81.75	81.80	82.04	85.63
	0 dB	77.95	79.68	90.09	83.34	85.39	88.17
	5 dB	78.78	80.54	94.98	85.56	86.30	90.00
	10 dB	80.21	81.77	96.36	89.08	88.29	90.04
	15 dB	82.58	75.44	95.93	88.57	89.26	92.12
	20 dB	81.88	75.26	95.93	90.68	89.77	93.68
AVE		79.58	78.28	**92.51**	86.51	86.84	89.94

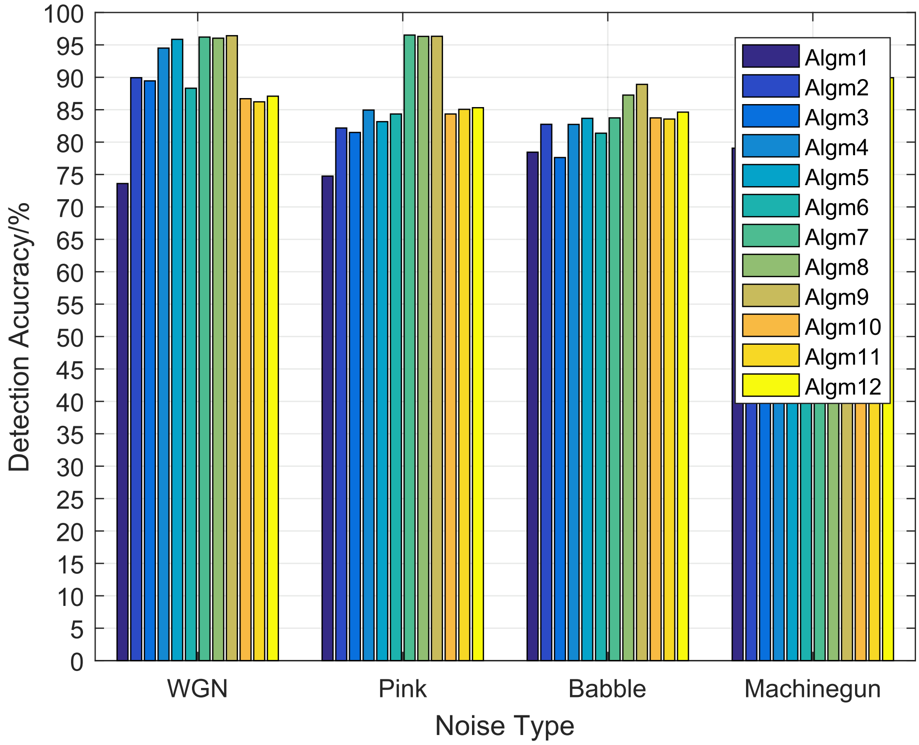


Fig. 9. The accuracies of different endpoint detection algorithms. Here *Algm1*: Threshold decision method based on short-time energy and ZCR; *Algm2*: Threshold decision method based on autocorrelation; *Algm3*: Threshold decision method based on energy-zero ratio; *Algm4*: Threshold decision method based on spectral variance; *Algm5*: Threshold decision method based on Sub-band energy entropy ratio; *Algm6*: Threshold decision method based on MFCC cepstrum distance; *Algm7*: Threshold decision method based on long-term signal variability; *Algm8*: Threshold decision method based on long-term spectral flatness measure; *Algm9*: Threshold decision method based on long-term spectral variability measure; *Algm10*: Decision method based on BP; *Algm11*: Decision method based on SVM. *Algm12*: Decision method based on DNN.

long-term methods and the neural network methods are the third category. It can be seen that in general, the long-term methods achieve the best performance compared with the other two categories in all noisy types. While the neural network category is in the second place (except for WGN conditions), and the short-time category performs the worst (except for WGN conditions).

Meanwhile, different algorithms can get different accuracies and performance as well. For the traditional short-time algorithms, the methods based on the short-time energy and ZCR, autocorrelation and the energy-zero ratio are in time-domain while the methods based on the spectral variance, sub-band energy entropy ratio and MFCC cepstrum distance are respectively in frequency domain, time–frequency and cepstrum domain. The detection accuracies of the algorithms in time domain are far lower than those of the detection algorithms in other domains in low SNR conditions. Among the four features in time domain, the detection accuracies based on short-time energy and ZCR are lower than those of the other two features, especially in low SNR environments. Moreover, the differences of the detection effect between the autocorrelation and the energy-zero ratio in different SNRs are not significant; compared with the features in time domain, spectral variance has stronger noise adaptability, especially in low SNRs (0, −5dB). However, it is of great difficulty to get high detection accuracies in low SNRs by using only one single feature. The appearance of the sub-band energy entropy ratio algorithm which combines the features both in time and frequency domain makes up for the shortcomings of the detection algorithms based on the features in single domain and thus achieving high detection accuracies. As for MFCC cepstrum distance, it has little difference with the algorithms in frequency domain in high SNRs, but under the low SNR condition (-5dB), it is not effective enough. Moving to the long-term algorithms, LSVM achieves the best performance among the three features except for PINK noise. For BP neural network and SVM, they are of similar performance for endpoint detection with indistinguishable experiment results in average. In addition, DNN is of the highest performance. In fact, the performance of the neural network fluctuates with the influence of different parameters and different training architectures, which is also the reason why we take the average data for comparison.

Besides the accuracies talked above, another two indexes, which need to be compared among these methods, are the database and the running time used in the whole process of the experiment. Actually, all the experiments in this paper are conducted on EBU database which has been mentioned in Section 5.2. As for the running time of the methods used in this paper, Fig. 10 gives a detailed data for comparison. Since the time used in the methods is random, and has few relationships with the noise conditions. Only among different categories, it is meaningful to compare the differences of the time used in each category. As shown in Fig. 10, the time used in the short-time methods and the long-term methods are averagely the same in general, and is obviously much less than the time used in training BP, SVM and DNN.

6. Discussion

The field has been exploring for more than one hundred years, and in context of the existing researches, the algorithm based on the neural network is in ascendant. However, designing a successful speech endpoint detection module still faces many challenges:

1) The current mainstream detection algorithms are basically for specific and common noise types. For more complex and non-stationary noise, there is no better detection algorithm yet.
2) When the feature of the speech information is extracted, it is very difficult to realize a high accuracy detection of the speech signal in low SNRs only by a single endpoint detection algorithm.
3) When the threshold method is used, the threshold values should be adaptable to the changes of the background noise. And the short-time impulse noise which exceeds the threshold, such as a person's smacking, a popping or a breathing, etc. should be marked as non– speech segments.
4) It can effectively smooth the inter-word gaps and eliminate the possible influence of the inter-word gaps on the endpoint detection.
5) For signals starting with the weak fricatives, bursts, nasals such as 'f', 't', 'k', 'n', etc, and the low-energy unvoiced speeches such as 'f', 'sh', 's', etc, which are similar to noise, the detection accuracy will decrease.
6) It can effectively handle the pauses between the syllables, and will not misjudge the temporary pauses between the syllables as the end of the speech.

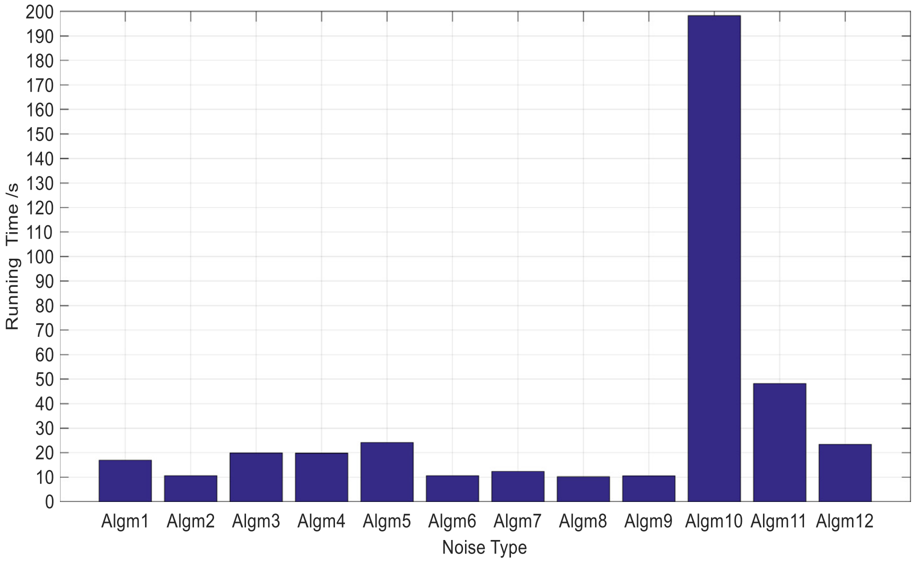


Fig. 10. The running time of different endpoint detection algorithms. Here *Algm1*: Threshold decision method based on short-time energy and ZCR; *Algm2*: Threshold decision method based on autocorrelation; *Algm3*: Threshold decision method based on energy-zero ratio; *Algm4*: Threshold decision method based on spectral variance; *Algm5*: Threshold decision method based on Sub-band energy entropy ratio; *Algm6*: Threshold decision method based on MFCC cepstrum distance; *Algm7*: Threshold decision method based on long-term signal variability; *Algm8*: Threshold decision method based on long-term spectral flatness measure; *Algm9*: Threshold decision method based on long-term spectral variability measure; *Algm10*: Decision method based on BP; *Algm11*: Decision method based on SVM. *Algm12*: Decision method based on DNN.

7) The detection mechanism can adaptively select an appropriate algorithm according to the characteristics of the dialects.
8) It's a challenge for us to take the characteristics of the signal phase into account, analyze the signal characteristics more comprehensively and then improve the endpoint detection accuracy.
9) The random vibration of speech endpoint detection may lead to the inaccuracy of the result detection.

7. Conclusion

This paper mainly analyzes the development course of the speech endpoint detection technology from the aspects of the time domain, frequency domain, cepstrum domain, and compares the detection accuracies of some classical algorithms.

Aiming at the problems existing in present algorithms, it is necessary for us to break through the traditional research framework and to combine the knowledge of interdisciplinary studies with the traditional algorithms. In addition, for threshold decision methods, setting up adaptable thresholds according to the characteristics of different speakers (including the sex, sound quality, emotion, etc.) in complex noise environments is one of the most important topics in the future. During the current mainstream pattern recognition algorithms, how to improve the detection accuracy combining with this sort of algorithm is also an important direction for the scholars.

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.

Acknowledgements

Tao Zhang provided the technical advice and some references. Yaqin Wu conceived and designed the study, and wrote the manuscript. Yangyang Shao and Wenli Du revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the Zhejiang Lab's International Talent Fund for Young Professionals.

References

[1] Lamel L, Rabiner L, Rosenberg AE, et al. An improved endpoint detector for isolated word recognition. J IEEE Trans Acoust Speech Signal Process 1981;29(4):777–85.
[2] Sakhnov K, Verteletskaya E, Simak B. Approach for energy-based voice detector with adaptive scaling factor. J Iaeng Int J Comput Sci 2009;36(4).
[3] Hussain A, Samad SA, Fah LB. Endpoint detection of speech signal using neural network Proceedings. IEEE. In: Presented at 2000 TENCON. p. 271–4.
[4] Liu HP, Li X, Xu BL, et al. A review and prospect of endpoint detection methods for speech signals. J Comput Appl Res 2008;25(8):2278–83.
[5] Rabiner LR, Sambur MR. An algorithm for determining the endpoints of isolated utterances. J Bell Syst Tech J 1957;54(2):297–315.
[6] Ali Z, Talha M. Innovative method for unsupervised voice activity detection and classification of audio segments. J IEEE Access 2018. 1 1.
[7] Katz MJ. Fractals and the analysis of waveforms. J Comput Biol Med 1988;18(3):145–56.
[8] Ghosh PK, Tsiartas A, Narayanan S. Robust voice ac-tivity detection using long-term signal variability. J Audio Speech Lang Process IEEE Trans 2011;19(3):600–13.
[9] Ma Y, Nishihara A. Efficient voice activity detection algorithm using long-term spectral flatness measure. J Eurasip J Audio Speech Music Process 2013;2013(1):1–18.
[10] Zhang Tao, Liu Yang, Ren Xiangying. Voice activity detection based on long-term spectral variability measure. J Front Comput Sci Technol 2019;2019(9):1534–42.
[11] Atal BS. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J Acoust Soc Am 1974;55(6):1304–22.
[12] Davis SB, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. Readings Speech Recognit 1990;28(4):65–74.
[13] Eshaghi M. MR Karami Mollaei, Voice activity detection based on using wavelet packet. Digital Signal Process 2010;20(4):1102–15.
[14] Li J, Zhou P, Jing X, et al. Speech endpoint detection method based on TEO in noise environment. Proc Eng 2012;29(4):2655–60.
[15] Chomorlig, Ze Z. Research on Endpoint Detection for Mongolian Speech Based on Support Vector Machine. In: 2011 International Conference on Intelligence Science and Information Engineering; 2011. pp. 290–94.
[16] Li L. Research of speech endpoint detection based on wavelet analysis and neural networks. J Electr Meas Instrum 2013.
[17] Sehgal A, Kehtarnavaz N. A Convolutional neural network smartphone App for real-time voice activity detection. IEEE Access 2018;99:1.
[18] Thomas S, Ganapathy S, Saon G, et al. Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions. In: Presented at 2014 IEEE Int. Conf. on acoustics, speech and signal processing. IEEE; 2014. p. 2519–23.
[19] Wilpon JG, Rabiner LR, Martin T. An improved word-detection algorithm for telephone-quality speech incorporating both syntactic and semantic constraints. Bell Labs Tech J 1984;63(3):479–98.
[20] Junqua JC, Reaves B, Mak B. A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognize. Presented at 1991 European Conference on Speech Communication and Technology, Eurospeech 1991, Genova, Italy, 1991.
[21] Craciun A, Gabrea M. Correlation coefficient-based voice activity detector algorithm. In: Presented at 2004 Canadian Conf. on Electrical and Computer Engineering. Canadian Conference on. IEEE. p. 1789–92.
[22] Jiqing Han, Speech Signal Processing, vol. 3, (School of Electronic Information, Wuhan University).
[23] Zhang RZ, Cui HJ. Speech endpoint detection algorithm analyses based on short-term energy. Audio Eng 2005;7:52–4.
[24] Petrou M, Kittler J. Optimal Edge Detectors for Ramp Edges. J IEEE Trans Pattern Anal Mach Intell 1991;13(5):483–91.
[25] Li Q, Zheng J, Tsai A, et al. Robust endpoint detection and energy normalization for real-time speech and speaker recognition. Speech Audio Process IEEE Trans 2002;10(3):146–57.
[26] Chen ZB, Xu B. Research on optimized speech endpoint detection algorithm based on Subband energy characteristics. J Acoust 2005;2:171–6.
[27] Li Q, Zheng J, Zhou Q, et al. Robust, real-time endpoint detector with energy normalization for ASR in adverse environments. In: Presented at 2001 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. Proceedings. IEEE; 2001. p. 233–6.
[28] Chen G, Liu J, Ye J. An improved method of endpoints detection based on energy-frequency-value. In: Presented at 2006 IEEE Conf. on High Density Microsystem Design and Packaging and Component Failure Analysis. IEEE; 2006. p. 9–11.
[29] Kumar S, Phadikar S, Majumder K. Modified segmentation algorithm based on Short Term Energy & Zero Crossing Rate for Maithili speech signal. In: Presented at 2017 IEEE Int. Conf. on Accessibility To Digital World. IEEE; 2017. p. 169–72.
[30] Jalil M, Butt FA, Malik A. Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals. In: Presented at 2013 IEEE Int. Conf. on Technological Advances in Electrical, Electronics and Computer Engineering. IEEE; 2013. p. 208–12.
[31] Shen JL, Hung JW, Lee LS. Robust entropy-based endpoint detection for speech recognition in noise environments. Presented at Int. Conf. on Spoken Language Processing, Incorporating the, Australian International Speech Science and Technology Conference, Convention Centre, Sydney, 1998.
[32] Jia C, Xu B. An improved entropy-based endpoint detection algorithm, 2002.
[33] Wu BF, Wang KC. Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments. IEEE Trans Speech Audio Process 2005;13(5):762–75.
[34] Wu GD, Lin CT. Word boundary detection with mel-scale frequency bank in noise environment. IEEE Trans Speech Audio Process 2000;8(5):541–54.
[35] Davis A, Nordholm S, Togneri R. Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold. Audio Speech Lang Process IEEE Trans 2006;14(2):412–24.
[36] Davis A, Nordholm S. A low complexity statistical voice activity detector with performance comparisons to ITU-T/ETSI voice activity detectors. In: Presented at Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. IEEE; 2004. p. 119–23.
[37] Wang Y, Qu BD, Li JB, et al. An improved endpoint detection algorithm based on band variance. Presented at 2007 Chinese academic annual conference on control and decision-making, 2007.
[38] Wang L, Cheng-Rong LI. An improved speech endpoint detection method based on adaptive band-partition spectral entropy. Comput Simul 2010.
[39] Guo Yu, Zhang Erhua, Liu Chi. An endpoint detection algorithm based on frequency-domain characteristics and transition fragment judgment. J Shandong Univ (Eng Sci) 2016;46(2):57–63.

[40] Zhang C, Dong M. An improved speech endpoint detection based on adaptive sub-band selection spectral variance. In: Presented at, IEEE Int Conf. on Control Conference. IEEE; 2013. p. 5033–7.

[41] Yang Xukui, Dan Qu, Zhang Wenlin, Yan Honggang. Adaptive Voice Activity Detection Based on Long-Term Information. Acta Electr 2018;46(4):878–85.

[42] Chuan-Yan WU, Fan YL. Speech endpoint detection based on speech time-frequency enhancement and spectral entropy. J Hangzhou Inst Electr Eng 2005;5:4682–4.

[43] Zhou MZ, Ji LX. Real-time endpoint detection algorithm combining time-frequency domain. In: Presented at 2010 IEEE Int. Conf. on International Workshop on Intelligent Systems and Applications. IEEE; 2010. p. 1–4.

[44] Haghani SK, Ahadi SM. Robust voice activity detection using feature combination. In: Presented at IEEE Int Conf. on Electrical Engineering. IEEE; 2013. p. 1–5.

[45] Wang Yinfeng, Huang Shaoguang, Wei Ying. A voice activity detection algorithm with sub-band detection based on time-frequency characteristics of mandarin. Presented at 2013 6th IEEE Int. Conf. on Image and Signal Processing. Hangzhou, China: IEEE; 2013.

[46] Morales-Cordovilla JA, Ma N, Sánchez V, et al. A pitch based noise estimation technique for robust speech recognition with missing data. In: Presented at 2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing. IEEE; 2011. p. 4808–11.

[47] Moradi N, Nasersharif B, Akbari A. Robust speech recognition using compression of Mel Sub-band energies and temporal filtering. In: Presented at IEEE Int Conf. on Telecommunications. IEEE; 2011. p. 760–4.

[48] Zhu C, Tian L, Li X, et al. Recognition of cough using features improved by Sub-band energy transformation. In: Presented at 2013 IEEE Int. Conf. on Biomedical Engineering and Informatics. IEEE; 2013. p. 251–5.

[49] Zhang Y, Wang K, Yan B. Speech endpoint detection algorithm with low signal-to-noise based on improved conventional spectral entropy. In: Presented at 2016 IEEE Int. Conf. on Intelligent Control and Automation. IEEE; 2016. p. 3307–11.

[50] Zaw Thein Htay, War Nu. The combination of spectral entropy, zero crossing rate, short-time energy and linear prediction error for voice activity detection. In: Presented at 2017 IEEE Int. Conf. on Computer and Information Technology. IEEE; 2017. p. 1–5.

[51] Haigh JA, Mason JS. Robust voice activity detection using cepstral features. In: IEEE Int. Conf. on IEEE Region 10 Conference on Tencon 93 Computer. IEEE; 1993. p. 321–4.

[52] Farzan A, Nourmohammadi A, Mashohor SB, et al. Novel voice activity detection based on Cepstrum moments. In: Presented at 2010 IEEE Int. Conf. on Computer and Automation Engineering. IEEE; 2010. p. 768–70.

[53] Wang H, Xu Y, Li M. Study on the MFCC similarity-based voice activity detection algorithm. In: Presented at 2011 IEEE Int. Conf. on Artificial Intelligence, Management Science and Electronic Commerce. IEEE; 2011. p. 4391–4.

[54] Chen Z, Weilan WU, Liu J, et al. Voice activity detection algorithm based on Mel cepstrum distance order statistics filter. J Univ Chin Acad Sci 2014;31 (4):524–9.

[55] Tao Z, Xiaobing Z, Mingxing Z. Speech endpoint detection with low SNR based on improved cepstrum distance method. J Audio Eng 2017.

[56] Cao Y, Dongsheng L, Jia S, et al. A speech endpoint detection algorithm based on wavelet transforms. In: Presented at 2014 Conf. Chinese Control and Decision Conference. p. 3010–2.

[57] Chen SH, Guido RC, Chen SH. Voice activity detection in car environment using support vector machine and wavelet transform. In: Presented at IEEE Int. Conf. IEEE International Symposium on Multimedia Workshops. IEEE; 2007. p. 252–5.

[58] Wu J, Zhang XL. Efficient multiple kernel support vector machine based voice activity detection. IEEE Signal Process Lett 2011;18(8):466–9.

[59] Huang NE, Shen Z, Long SR, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc Math Phys Eng Sci 1971;1998(454):903–95.

[60] Li MM, Yang HW, Hong N, et al. Endpoint detection based on EMD in noise environment. In: Presented at 2011 IEEE Int. Conf. on Computer Sciences and Convergence Information Technology. IEEE; 2011. p. 783–787****.

[61] Kim SK, Kang SI, Park YJ, et al. Power spectral deviation-based voice activity detection incorporating teager energy for speech enhancement. Symmetry-Basel 2016;8(7):58.

[62] Feng C, Zhao C. Voice activity detection based on ensemble empirical mode decomposition and teager kurtosis. In: Presented at 2015 IEEE Int. Conf. on Signal Processing. IEEE; 2015. p. 455–60.

[63] Wu J, Zhang XL. An efficient voice activity detection algorithm by combining statistical model and energy detection. EURASIP J Adv Signal Process 2011;2011(1):18.

[64] Bergh TF, Hafizovic I, Holm S. Multi-speaker voice activity detection using a camera-assisted microphone array. In: Presented at 2016 IEEE Int. Conf. on Systems, Signals and Image Processing. IEEE; 2016. p. 1–4.

[65] Li X, Horaud R, Girin L, et al. Voice activity detection based on statistical likelihood ratio with adaptive thresholding. In: Presented at 2016 IEEE Int. Conf. IEEE International Workshop on Acoustic Signal Enhancement. IEEE; 2016. p. 1–5.

[66] Obuchi Y. Framewise speech-nonspeech classification by neural networks for voice activity detection with statistical noise suppression. In: Presented at 2016 IEEE Int. Conf. on Acoustics, Speech and Signal Processing. IEEE; 2016. p. 5715–9.

[67] Li J, You D. Enhanced Speech Based Jointly Statistical Probability Distribution Function for Voice Activity Detection. J Chinese Journal of Electronics 2017;26 (2):325–30.

[68] Ng T, Zhang B, Long N, et al. Developing a Speech Activity Detection System for the DARPA RATS Program. 2012.

[69] Eyben F, Weninger F, Squartini S, et al. Real-life voice activity detection with LSTM Recurrent Neural Networks and an application to Hollywood movies. Presented at 2015 IEEE Int. Conf. on Acoustics, Speech and Signal Processing. IEEE, 2013, 483-487.

[70] Boonkla S, Sertsi P, Chunwijitra V, et al. Robust Voice Activity Detection Based on LSTM Recurrent Neural Networks and Modulation Spectrum. Presented at Asian Conf. Asia-Pacific Signal and Information Processing Association. IEEE, 2017.

[71] Zi-Chen Fan, Zhongxin Bai, Xiao-Lei Zhang, et al. AUC Optimization for Deep Learning Based Voice Activity Detection. J ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[72] Zhang XL, Wang D. Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection. Cse Ohio 2014.

[73] Chen J, Wang Y, Wang DL. A feature study for classification-based speech separation at low signal-to-noise ratios. IEEE/ACM Trans Audio Speech Lang Process 2014;22(12):1993–2002.

[74] Zhang XL, Wang DL. Boosting Contextual Information for Deep Neural Network Based Voice Activity Detection. IEEE/ACM Trans. Audio Speech & Language Processing. 2016;24(2):252–64.

[75] Kim J, Hahn M. Voice Activity Detection Using an Adaptive Context Attention Model. J IEEE Signal Processing Letters 2018. 1 1.

[76] L Wang, K Phapatanaburi, Z Go, et al, Phase aware deep neural network for noise robust voice activity detection. Presented at 2017 IEEE Int. Conf. on Multimedia and Expo. IEEE Computer Society, 1087-1092.

[77] Shahsavari S, Sameti H, Hadian H. Speech activity detection using deep neural networks. IEEE: C Electrical Engineering; 2017.