Speech endpoint detection in fixed differential beamforming combined with modulation domain

Hongli Wang¹, Qingning Zeng² and Xuejun Zhao¹*

¹ College of Information and Communication, Guilin University of Electronic Technology, Guilin, Guangxi, 541004, China
² College of Information and Communication, Guilin University of Electronic Technology, Guilin, Guangxi, 541004, China
*Corresponding author’s e-mail: 19022201034@mails.guet.edu.cn

Abstract. The actual speech signal is always in a complex environment, which affects the accuracy of speech endpoint detection. In order to improve the accuracy of speech endpoint detection under low Signal-To-Noise Ratio(SNR), a speech endpoint detection technology based on fixed differential beamforming and modulation domain is proposed. Firstly, two microphones of the first-order differential microphone arrays(DMAs) are used to collect the target speech and noise, and then a fixed differential beamformer is used to reduce the noise. Then, the phase compensated modulation domain spectral subtraction method is used to further eliminate the residual noise. Finally, the subband spectral entropy algorithm is used to detect the speech signal endpoint. The experimental results show that, compared with the existing methods, the proposed method has a higher accuracy of speech endpoint detection in the low SNR environment of -10~5dB.

1. Introduction
Speech endpoint detection is to divide a given speech signal into the part containing speech and the other part containing only noise and interference [1]. Accurate detection of the presence or absence of speech is challenging, especially when the speech signal is destroyed by background noise [2].

At present, there are some classical endpoint detection methods, such as spectral entropy method [3], wavelet transform method [4], cepstrum distance [5]. These classical endpoint detection methods, most of the SNR is at 10dB, endpoint detection performance is good. When the SNR is reduced, the quality of endpoint detection is significantly reduced [6], resulting in false detection of speech.

Therefore, the endpoint detection under low SNR is concerned. Parvin Ahmadi et al. [7] proposed a new speech endpoint detection method based on sparse representation, which has good performance under the condition of low SNR and is superior to existing speech endpoint detection methods. In-Chul Yoo et al. [8] proposed a robust speech activity detection algorithm based on formants, which overcame the problem of formants detection under severe noise conditions. In various noise environments, the processing time of this method is greatly reduced, and it has strong robustness to various noises, and is suitable for various applications. Reference [9] combined the improved spectral subtraction algorithm with the energy-entropy ratio, and achieved the correct endpoint detection target through noise reduction under the condition of low SNR. Reference [10] proposed an endpoint detection algorithm based on modulation domain spectral subtraction combined with autocorrelation function. When the speech is in low SNR, the position of the speech endpoint can be detected better.
However, the accuracy of endpoint detection still needs to be improved. Reference [11] combines spectral subtraction with uniform subband frequency band variance algorithm, but the use of spectral subtraction for speech enhancement will produce music noise and reduce endpoint detection performance.

In order to improve the accuracy of endpoint detection in low SNR, a speech endpoint detection technology based on fixed differential beamforming combined with modulation domain is proposed in this paper. Modulation domain spectrum subtraction of fixed differential beamforming and phase compensation is performed for the acoustic signals collected from the first-order DMAs. Subband spectrum entropy method is used for endpoint detection. Simulation results show that the proposed algorithm can achieve better endpoint detection effect when the SNR is low.

2. Algorithm framework in this paper

Fig 1 is the principle block diagram of the algorithm in this paper, where \(x_1\) and \(x_2\) are the speech signals.

Collected by mic_1 and mic_2 of the first-order DMAs respectively. After the fixed difference beamforming, the speech signals obtained by the two microphones are output as \(y_0\), and then the stationary noise is suppressed by the modulation domain spectrum subtraction method with phase compensation to obtain the output signal \(y_1\). Finally, the subband spectrum entropy algorithm is adopted to carry out endpoint detection.

![Figure 1. Schematic diagram of the proposed algorithm](image)

3. Speech enhancement

3.1. Microphone array

Microphone array uses the spatial direction information of speech and interference to enhance the speech signal in a certain target direction, which can be divided into two methods: fixed beamforming and adaptive beamforming [12]. In this paper, fixed differential beamforming is used to enhance speech in time domain.

Fig 2 shows the first-order DMAs, consisting of mic_1 and mic_2. The two microphones are in the same line, forming a delayed subtraction beamformer.

Let the sampling frequency of the system be \(f_s\), the sound velocity is \(c\), and the center distance between the two microphones is \(d\). Its calculation formula is as follows:

\[
d = \frac{c}{f_s}
\]  

(1)

As shown in Fig 2, if the sound source is located on the two microphone lines and at one end of mic_1, and the noise source is located on the two microphone lines and at one end of mic_2, the noisy speech obtained by mic_1 and the noisy speech obtained by mic_2 are differential calculated, as shown in Equation (2):

\[
y_1(i) = x_1(i) - x_2(i-1)
\]  

(2)

Where \(x_1\) represents the noisy speech obtained by mic_1, and \(x_2\) represents the noisy speech obtained by mic_2.

The subtraction of mic_1 signal and mic_2 signal keeps the desired direction signal and forms a zero trap in the undesired direction, thus achieving beamforming in the fixed source direction. This method can suppress the signal in the direction of the fixed noise source, and effectively suppress both coherent and incoherent noises, and enhance the speech.
In the difference operation, the noise signal is eliminated and the speech signal is retained, but the speech signal is distorted, which needs to be filtered and restored by FIR recovery filter, as shown in Equation (3):

\[ y_2(i) = \sum_{k=0}^{L} h(k) y_1(i-k) \]  

(3)

\[ h = [h(0) \ h(1) \ h(2) \ldots h(L)] \]  

(4)

Where \( h \) is the filter coefficient, is a \((L-1)\) dimensional row vector; \( y_1 \) is the output signal of fixed differential beamforming.

In fact, the noise is not located in the connection between the two microphones and at one end of mic_2. Therefore, the noise cannot be completely eliminated by differential operation. The residual noise gain can be determined by the heart-shaped pointing beam diagram of the first-order DMAs, as shown in Fig3. Noise coverage 360°, the closer the noise is to the target speech \((0°)\) direction, the smaller the attenuation of noise; in the direction \(180°\), a zero-trap area is formed, and the noise attenuation is maximum.

3.1.1 Experiment Analysis. To evaluate the performance of the propose localization system, a series of simulation experiments were conducted. In this study, the room size is set 7m, 4m and 3m for the length, width and height, respectively. The first-order DMAs is located in the middle of the laboratory. The target speech is on the connection between the two microphones and is located at 1m directly in front of mic_1. The noise is located at the direction of \(90°, 150°, 180°\) and the distance from mic_2 is 1m. White Gaussian noise (white) and babble noise from Noisex-92 noise library are selected in the experiment, and the sampling frequency is set at 16khz.

Noise reduction under different noise environments is shown in Fig4(a) and (b) are noise reduction comparison figures of different angles under white noise environment and babble noise environment. The same conclusion can be obtained, that is, the maximum noise reduction is achieved when the noise is located in the direction \(180°\).

Fig5(a) shows two channels of clean speech collected by the first-order DMAs; white noise of -10dB was added to the two channels of collected clean speech, as shown in Fig5(b). Fig5 (c) shows the time domain waveform of the output signal of the fixed difference beamforming, with the SNR improved by 8.3dB. Although the difference operation suppressed the noise to some extent, the noise still existed. In order to further eliminate the residual noise and make the endpoint detection more accurate, the modulation domain spectrum subtraction with phase compensation was introduced in this paper for further noise reduction.
3.2. Phase compensated modulation domain spectral subtraction

3.2.1. Modulation domain. In actual operation, it is assumed to $y(n)$ be a noise-polluted input signal, which is composed of clean speech $x(n)$ and additive noise $d(n)$, as shown in Equation (5):
\[ y(n) = x(n) + d(n) \]  

(5)

The short-time Fourier transform of Equation (5) is shown in Equation (6):

\[ Y(n, k) = \sum_{l=-\infty}^{\infty} y(l) w(n-l) e^{-\frac{j 2\pi kl}{N}} \]  

(6)

Where \( k \) is the discrete frequency variable, \( N \) is the frame length, \( w(n) \) represents the window function. The polar coordinate form of Equation (6) is shown in Equation (7):

\[ Y(n, k) = \left| Y(n, k) \right| e^{j\angle Y(n, k)} \]  

(7)

Where \( \left| Y(n, k) \right| \) and \( \angle Y(n, k) \) represent the amplitude spectrum and phase spectrum of speech respectively. Then the polar form of the modulation domain can be obtained

\[ Y(\zeta, n, k) = \left| Y(\zeta, n, k) \right| e^{j\angle Y(\zeta, n, k)} \]  

(8)

Where \( \zeta \) is the modulation frame variable, \( k \) is the discrete frequency variable, \( n \) is the discrete modulation frequency variable, \( \left| Y(\zeta, n, k) \right| \) and \( \angle Y(\zeta, n, k) \) is the amplitude spectrum and phase spectrum of the modulation domain respectively.

3.2.2. Modulation domain spectrum subtraction. In the modulation domain, spectral subtraction can be used to improve the amplitude of the modulation domain [13], as shown in Equation (9):

\[
Z(\zeta, \zeta, m) = \begin{cases} 
\left( \left| Y(\zeta, k, m) \right| - \mu |\hat{D}(\zeta, k, m)| \right)^\frac{1}{\gamma} & \text{if } \left| Y(\zeta, k, m) \right| \geq (\mu + \delta) |\hat{D}(\zeta, k, m)| \\
(\delta)^\frac{1}{\gamma} & \text{others}
\end{cases}
\]  

(9)

Where \( \mu \) is the overreduction factor, \( \delta \) is the inhibitory factor, \( \gamma \) determines the spectrum reduction type, if \( \gamma = 1 \) is the amplitude spectrum reduction, \( \gamma = 2 \) is the energy spectrum reduction. \( \hat{D}(\zeta, k, m) \) is the modulation spectrum estimation of noise, which can be updated by activity detection (VAD). As shown in Equation (10):

\[ |\hat{D}(\zeta, k, m)|^\gamma = l |\hat{D}(\zeta - 1, k, m)|^\gamma + (l-1) |Y(\zeta, k, m)|^\gamma \]  

(10)

Where \( l \) is the forgetting factors, when the signal is a noise segment, the noise estimation is updated.

3.2.3. Modulation domain phase compensation. Since the noisy signal is a real signal, the modulation spectrum obtained through STFT is conjugate symmetric [14], as shown in Equation (11):

\[ Y(\zeta, k, m) = Y^*(\zeta, N-k, m) \]  

(11)

An antisymmetric function can be used to compensate the modulation phase spectrum, and the phase compensation function is expressed in Equation (12):

\[ \Lambda(\zeta, k, m) = e^{\phi(m)} |\hat{D}(\zeta, k, m)| \]  

(12)
Where the phase compensation degree $\Lambda(\zeta,k,m)$ is calculated from the modulation amplitude spectrum of the noise $\hat{D}(\zeta,k,m)$, and $\epsilon$ is an empirical constant and $\varphi(m)$ is antisymmetric function, as shown in Equation (13):

$$\varphi(m) = \begin{cases} 
1, & 0 < \frac{m}{NIS} < 0.5 \\
-1, & 0.5 < \frac{m}{NIS} < 1 \\
0, & \text{other} 
\end{cases}$$  \hspace{1cm} (13)

Where $NIS$ is the number of frames of the leading no-talk segment. The improved modulation spectrum $Z_A(\zeta,k,m)$ is obtained by adding the modified amplitude estimate of the modulation domain $\hat{Y}(\zeta,k,m)$ and the phase compensation degree $\Lambda(\zeta,k,m)$, as shown in Equation (14):

$$Z_A(\zeta,k,m) = \hat{Y}(\zeta,k,m) + \Lambda(\zeta,k,m)$$  \hspace{1cm} (14)

Then the improved modulation phase spectrum can be expressed as:

$$\angle Z_A(\zeta,k,m) = \text{arg}[Z_A(\zeta,k,m)]$$  \hspace{1cm} (15)

Where $\text{arg}$ represents the Angle of the complex number.

Finally, the improved modulation phase spectrum and amplitude spectrum are combined to obtain the final improved modulation spectrum, as shown in Equation (16):

$$Z(\zeta,k,m) = |Z(\zeta,k,m)| e^{i\angle Z_A(\zeta,k,m)}$$  \hspace{1cm} (16)

In order to verify the algorithm performance of fixed differential beamforming combined with modulated domain spectral subduction. White, babble, automobile (volvo) and factory noises from Noisex-92 noise library are selected in the experiment and the noises are located in the direction $180^\circ$. The SNR was set as 5dB, 0dB, -5dB and -10dB. Simulation experiments verify the performance of the proposed speech enhancement algorithm from the aspects of speech time domain waveform contrast and SNR total noise reduction. The reference algorithms are Logarithm Minimum Mean Square Error(LogMMSE), Spectral Subtraction(SS) and Improved Wiener Filter(IWF) algorithm respectively.

Fig6 shows that the comparison of different speech enhancement algorithms. Fig6(a) shows that the waveform of clean speech signal, and Fig6(b) shows the waveform of speech with noise in the white noise environment with a SNR of -10dB. Fig6(d) shows that after IWF processing, there is a large of residual noise and speech distortion. Fig6(e) shows that SS introduces "music noise" to cause speech distortion. Fig6(f) shows that the noise elimination effect of fixed difference beamforming combined with modulated domain spectrum reduction is improved to a certain extent, with less noise residue.

In table1, compared with IWF, LogMMSE and SS, the proposed method has the better performance and the SNR has been greatly improved. Furthermore, the residual noise is small, which provides good data for end-point detection in the later stage.
Table 1. The SNR enhancement comparison of each algorithm under different environments

| SNR  | Noise | IWF   | LogMMSE | SS     | Proposed Method |
|------|-------|-------|---------|--------|-----------------|
| 5dB  | white | 11.3072 | 13.4510  | 16.7049 | 34.2206         |
|      | volvo | 12.3952 | 13.0017  | 16.6765 | 22.9629         |
|      | factory | 10.5983 | 13.5101  | 11.3923 | 22.5774         |
|      | babble | 7.9897  | 14.1630  | 6.1037  | 31.2063         |
| 0dB  | white | 11.2418 | 13.4509  | 15.4278 | 33.9166         |
|      | volvo | 12.3343 | 12.9993  | 15.6808 | 22.5015         |
|      | factory | 11.8172 | 13.5101  | 11.3560 | 18.4200         |
|      | babble | 8.5229  | 14.1632  | 5.4976  | 30.5370         |
| -5dB | white | 11.2075 | 13.4508  | 12.0273 | 31.6651         |
|      | volvo | 12.3086 | 12.9980  | 11.1566 | 19.7531         |
|      | factory | 11.7637 | 13.5100  | 10.9985 | 16.1922         |
|      | babble | 8.4684  | 14.1633  | 3.9564  | 29.5988         |
| -10dB| white | 11.0084 | 13.4508  | 9.0809  | 28.9364         |
|      | volvo | 12.3000 | 12.9972  | 11.0846 | 14.4216         |
|      | factory | 11.7470 | 13.5100  | 11.0343 | 16.2893         |
|      | babble | 8.4496  | 14.1633  | 3.4584  | 27.6791         |

4. Endpoint detection of subband spectral entropy

The idea of subband spectral entropy is to divide a frame into several subbands and then calculate the spectral entropy of each subband. So as to eliminate the influence of noise on the amplitude of each spectral line.

Set the time domain waveform of voice signal containing noise is \( x(n) \), and the first voice signal obtained after window frame processing is \( x_i(m) \), and its DFT is

\[
X_i(k) = \sum_{m=0}^{N-1} x_i(m) \exp(-i2\pi km / N)
\]  

(17)

Where \( X_i(k) \) is the short-time Fourier transform of the speech frame \( x_i(m) \), and the energy of each component \( Y(k) = |X_i(k)|^2 \). In this case, the probability density function with energy is defined as:
The information entropy is calculated for the positive frequency part of each frame

\[ H(i) = -\sum_{k=0}^{N/2} p(k,i) \log p(k,i) \]  

(19)

Where \( H(i) \) is the spectral entropy of frame \( i \).

Each frame is divided into several subbands. Let each subband be composed of 4 spectral lines and have \( N_b \) subbands in total. In this way, the energy of subband \( m \) in frame \( i \) is

\[ E_b(m,i) = \sum_{k=(m-1)*4}^{(m-1)*4+3} Y_i(k), 1 \leq m \leq N_b \]  

(20)

Correspondingly, the probability \( p_b(m,i) \) of subband energy is:

\[ p_b(m,i) = \frac{E_b(m,i)}{\sum_{k=1}^{N_b} E_b(m,i)}, 1 \leq m \leq N_b \]  

(21)

The spectral entropy of subband \( H_b(i) \) is:

\[ H_b(i) = -\sum_{m=1}^{N_b} p_b(m,i) \log p_b(m,i) \]  

(22)

In the calculation of spectral entropy, a normal quantity \( K \) is introduced into the probability distribution (22), and a new probability density formula of subband energy is obtained

\[ p_b'(m,i) = \frac{E_b(m,i)}{\sum_{k=1}^{N_b} (E(k,i) + K)}, K > 0 \]  

(23)

The new subband spectral entropy \( H_b'(i) \) is:

\[ H_b'(i) = -\sum_{m=1}^{N_b} p_b'(m,i) \log p_b'(m,i) \]  

(24)

In the noise environment, the distinction between speech signal and noise signal can be improved by introducing the normal quantity \( K \).

5. Simulation results and analysis

The experimental acquisition equipment uses M-Track Eight collector of M-Audio Company to collect speech data, and the pickup is Omni-directional miniature microphone. The software used for voice acquisition is Cubase, an advanced audio acquisition software of Steinberg Company, which can select multiple data acquisition and processing. The recording tool is the first-order DMAs (see Figure 2). The position of the target speech is on the lines between the two microphones and located at 1m directly in front of mic_1, and the noise is located at 120°, 150° and 180°. In order to reduce the length of this paper, the speech endpoint detection results when the noise is located 180° are
emphatically described. Set the sampling frequency \( f_s \) is 16kHz and the sound velocity \( c \) is 340m/s. According to Equation(1), the center distance between the two microphones \( d \) is 2.125cm. The recording environment was relatively quiet in an empty conference room. Clean speech and noise signals were played through high-fidelity speakers. The simulation tool was MATLAB 2016a.

Four kinds of background noises are used in the experiment, namely, white, volvo, factory and babble. The above four kinds of background noises are all from Noisex-92 noise library. The clean speech was verified with the noisy speech with the SNR of 5dB, 0dB, -5dB and -10dB respectively. Fig7~10 and Table 2 are the results of noise under the condition 180°.

The comparison algorithm in this paper are short-time energy, modulation domain spectrum subtraction combined with autocorrelation function and spectrum subtraction combined with variance method. The three algorithms are compared with the proposed algorithm under different noises and different SNR. For convenience, the latter two algorithms are respectively abbreviated as reference10 and reference 11. Fig7–10 are the verification results of the four algorithms in the Gaussian white noise environment with a SNR of -10dB, and (a) is all clean speech signal waveforms. (b) is the noise of white Gaussian noise with a SNR of -10dB. (c) is detect waveforms for the speech endpoints of each algorithm. The vertical solid line is the starting point of speech, and the vertical dotted line is the end point of speech.

![Figure 7. Short-time energy](image1)

![Figure 8. Reference 10](image2)

![Figure 9. Reference 11](image3)

![Figure 10. Propose Method](image4)

Fig7(c) shows the speech endpoint detection results of the short-time energy algorithm. Since both speech and noise have energy and are very sensitive to the signal amplitude, the speech is completely submerged in the noise in the low SNR environment and loses the ability of endpoint detection. Fig8 shows the endpoint detection results of reference 10. In a low SNR environment, because a large number of noise signals remain after the modulation domain spectrum reduction and speech distortion.
The endpoint detection results of Fig8(c) shows that the amplitude of autocorrelation function between 1.6s and 2.2s of speech is too small, and the speech is mistakenly detected as noise. As shown in Fig9, the endpoint detection results in Reference 11 show that although the multi-window spectrum estimation spectrum reduction effectively reduces the noise in the contaminated signal, part of the speech information is lost. Fig9(c) shows that short-term energy entropy ratio is larger at the beginning of speech, but the amplitude is too small after 1.5s, resulting in speech missed detection. Fig10(c) shows the algorithm in this paper, which has good detection ability and robustness. Compared with Fig7-9, the speech endpoint detection has a higher accuracy rate.

In order to make a more intuitive comparison of the actual performance of each algorithm, Table 2 shows the comparison of the accuracy rates of the four detection algorithms under the four noise environments of white, factory, volvo and babble. Accuracy can be calculated by the following formula:

$$\frac{\text{Total Frames} - \text{Error Frames}}{\text{Total Frames}} \times 100\%$$

Tab2 The accuracy comparison of each algorithm in different environments

| SNR   | Noise | short-time energy | Reference10 | Reference11 | Proposed Method |
|-------|-------|-------------------|-------------|-------------|-----------------|
| 5dB   | white | 86.4%             | 92.7%       | 90.8%       | 96.7%           |
|       | volvo | 82.3%             | 94.9%       | 93.4%       | 95.8%           |
|       | factory | 83.2%       | 89.7%       | 86.0%       | 92.0%           |
|       | babble | 81.3%           | 92.2%       | 83.2%       | 95.0%           |
| 0dB   | white | 78.4%             | 91.6%       | 84.4%       | 94.6%           |
|       | volvo | 78.0%             | 93.0%       | 90.1%       | 94.5%           |
|       | factory | 80.1%       | 86.3%       | 80.7%       | 89.8%           |
|       | babble | 82.0%           | 91.2%       | 79.0%       | 93.6%           |
| -5dB  | white | 78.2%             | 88.0%       | 79.0%       | 93.4%           |
|       | volvo | 71.6%             | 90.8%       | 88.6%       | 92.6%           |
|       | factory | 74.5%       | 82.0%       | 77.0%       | 84.3%           |
|       | babble | 72.9%           | 87.3%       | 74.0%       | 93.3%           |
| -10dB | white | 57.6%             | 78.2%       | 69.0%       | 89.3%           |
|       | volvo | 47.5%             | 84.7%       | 85.2%       | 89.1%           |
|       | factory | 61.1%       | 72.6%       | 67.2%       | 77.2%           |
|       | babble | 37.5%           | 84.7%       | 68.0%       | 88.4%           |

As can be seen from Table 2, in the noise environment of 5dB and 0dB, the accuracy rate of the four algorithms is above 70%, with a good speech recognition rate. However, in the noise environment of -5dB and -10dB, the accuracy of the four algorithms decreases to a certain extent. Among them, the detection performance of short-time energy decreases most obviously. In the -10dB babble noise environment, the accuracy rate is only 37.5%, which shows a serious misjudgment phenomenon. By contrast, the proposed method can obtain relatively high accuracy under low SNR. Accuracy rate of
77.2% in the factory noise environment of -10dB and the highest accuracy rate of 93.4% in the white noise environment of -5dB. Obviously, the algorithm in this paper still maintains good robustness and robustness even under low SNR, and can accurately judge the endpoint. Result revealed this fact that the proposed algorithm has better feasibility.

Table 3 and Table 4 respectively show the endpoint detection accuracy of the proposed algorithm in the white and babble noise environment and when the noise is located at 120° and 150°. By comparison with Table (2), the following conclusions can be drawn: 1) The accuracy of proposed method is higher when the noise is located 180° than 120° and 150°; 2) When the noise is white and located at 120° and 150°, the endpoint detection accuracy is still higher than that of the contrast algorithm; 3) When the noise is babble and located at 120° and 150°, the accuracy of endpoint detection is higher than that of short-term energy and reference 11, but slightly lower than that of reference 10.

### Tab3 The accuracy of this algorithm in white noise environment

| Noise angle | SNR  | Proposed Method |
|-------------|------|-----------------|
| white 120°  | 5dB  | 0.94            |
|            | 0dB  | 0.92            |
|            | -5dB | 0.93            |
|            | -10dB| 0.81            |
|            | 5dB  | 0.96            |
| 150°        | 0dB  | 0.92            |
|            | -5dB | 0.91            |
|            | -10dB| 0.87            |

### Tab4 The accuracy of this algorithm babble noise environment

| Noise angle | SNR  | Proposed Method |
|-------------|------|-----------------|
| babble 120° | 5dB  | 0.89            |
|            | 0dB  | 0.84            |
|            | -5dB | 0.80            |
|            | -10dB| 0.79            |
|            | 5dB  | 0.93            |
| 150°        | 0dB  | 0.95            |
|            | -5dB | 0.92            |
|            | -10dB| 0.85            |

### 6. Summary

This paper proposes a joint fixed difference beamforming modulation domain spectrum subtraction speech endpoint detection technology. The stationary and non-stationary noise is eliminated by the fixed differential beamforming operation of the two channels of speech signal collected by the first-order differential microphone array. Then the modulation domain spectral subtraction is used for noise reduction. Finally, the subband spectral entropy algorithm is used to detect the endpoints. Experimental results show that proposed method can effectively improve the accuracy of endpoint detection under low SNR. It not only reduce the false detection of speech, achieve more ideal detection results, and has a strong adaptability. The next step is to use the dual microphone array for speech endpoint detection.
Fund of the project
National Natural Science Foundation of China (61961009); Key Project of Guangxi Natural Science Foundation(2016GXNSFDA380018); Foundation of Guangxi Wireless Broadband Communication and Signal Processing Key Laboratory(GXKL06200107)

References
[1] Ariav I, Cohen I. (2019) An End-to-End Multimodal Voice Activity Detection Using WaveNet Encoder and Residual Networks. IEEE Journal of Selected Topics in Signal Processing, 13: 265-274.
[2] Soares A S P, Parreira W D, Souza E G, et al.(2018) Energy-based voice activity detection algorithm using Gaussian and Cauchy kernels. In: 2018 IEEE 9th Latin American Symposium on Circuits & Systems (LASCAS). Puerto Vallarta. pp: 1-4.
[3] Zhao H, Zhao L, Zhao K, et al. (2019) Voice Activity Detection Based on Distance Entropy in Noisy Environment. In: 2009 Fifth International Joint Conference on INC, IMS and IDC. Seoul, Korea (South). pp: 1364-1367.
[4] Yali C, Dongsheng L, Shuo J, et al. (2014) A speech endpoint detection algorithm based on wavelet transforms. In: The 26th Chinese Control and Decision Conference (2014 CCDC). Changsha, china. pp: 3010-3012.
[5] Haigh J A, Mason J S. (1993) Robust voice activity detection using cepstral features. In: Proceedings of TENCON '93. IEEE Region 10 International Conference on Computers, Communications and Automation. Beijing, China. pp: 321-324.
[6] Song Z.Y. (2013) Application of MATLAB in Speech Signal and Synthesis. Beijing University of Aeronautics and Astronautics Press. Beijing.
[7] Ahmadi P Joneidi M. (2014) A new method for voice activity detection based on sparse representation. In: 2014 7th International Congress on Image and Signal Processing. Dalian. pp: 878-882.
[8] Yoo I C, Lim L, Youk D. (2015) Formant-Based Robust Voice Activity Detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23: 2238-2245.
[9] Zhao F. (2015) Speech Endpoint Detection Algorithm Based on Spectral Subtraction of Multi-window Spectrum Estimationand Energy Entropy Ration. Journal of Chaohu University, 18:80-85.
[10] Wang Y, Zeng Q.N, Long C, et al. (2018) An improved method for speech endpoint detection in low SNR environment. Technical Acoustics, 37: 457-467.
[11] Wang W, Hu G.M, Yang L, et al. (2016) Endpoint detection based on spectral subtraction method and uniform subband frequency band variance method. Audio Engineering, 40: 40-43+66.
[12] Yuan K, Liang W.Q, Zhang G.X. (2009) Speech Enhancement Based on Equilateral Triangle Microphone Array. Journal of Tsinghua University (Natural Science Edition) Network, 49: 971-974.
[13] Chen Z.Q, Li X.Y, Xie Y.L. (2015) Modulation Domain Spectrum Subtraction Combined with Phase Spectrum Compensation. Journal of Signal Processing, 31:468-473.
[14] Bu Y.T, Zeng Q.N, Zheng Z.H. (2020)A Speech Endpoint Detection Algorithm in Low SNR Environment. Technical Acoustics, 39:592-602.