Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 3608286, 19 pages
https://doi.org/10.1155/2020/3608286

Research Article

An Improved Speech Segmentation and Clustering Algorithm Based on SOM and *K*-Means

Nan Jiang[1] **and Ting Liu**[2]

[1]*Criminal Investigation Police University of China, Shenyang 110854, China*
[2]*Liaoning University, Shenyang 110036, China*

Correspondence should be addressed to Ting Liu; liuting_tinka@sina.cn

Received 14 May 2020; Accepted 27 July 2020; Published 12 September 2020

Academic Editor: Thomas Hanne

This paper studies the segmentation and clustering of speaker speech. In order to improve the accuracy of speech endpoint detection, the traditional double-threshold short-time average zero-crossing rate is replaced by a better spectrum centroid feature, and the local maxima of the statistical feature sequence histogram are used to select the threshold, and a new speech endpoint detection algorithm is proposed. Compared with the traditional double-threshold algorithm, it effectively improves the detection accuracy and antinoise in low SNR. The *k*-means algorithm of conventional clustering needs to give the number of clusters in advance and is greatly affected by the choice of initial cluster centers. At the same time, the self-organizing neural network algorithm converges slowly and cannot provide accurate clustering information. An improved *k*-means speaker clustering algorithm based on self-organizing neural network is proposed. The number of clusters is predicted by the winning situation of the competitive neurons in the trained network, and the weights of the neurons are used as the initial cluster centers of the *k*-means algorithm. The experimental results of multiperson mixed speech segmentation show that the proposed algorithm can effectively improve the accuracy of speech clustering and make up for the shortcomings of the *k*-means algorithm and self-organizing neural network algorithm.

1. Introduction

Speech segmentation is an essential basic work in speech recognition and speech synthesis, and its quality has a huge impact on the follow-up speech recognition. Although the manual segmentation and annotation have high accuracy, they are time-consuming and require skilled domain experts to complete, so automatic speech segmentation has become a hot research topic in speech processing [1].

The speaker segmentation and clustering in this paper is to segment the continuous audio containing the speech of several speakers into several speech segments, so that each speech segment only contains the speech of one speaker. Then, the speech segments of the same speaker are grouped together and marked with distinctive labels, which determines who is speaking and when. This task is also known as speaker diarization [2–4].

Speaker segmentation and clustering technology as an important front-end processing technology can get the speaker change information in the audio, which can facilitate the subsequent speech processing applications, such as speech recognition and further machine translation, and grammar analysis. Most of the current voice processing technology is for a single [5–7], etc. only for audio files containing a person speaking, but when the audio contains multiple people speaking, it cannot meet the demand. At present, the speaker segmentation and clustering system has achieved good performance on two-person telephone conversation data, but it still faces many challenges in complex scenes such as conference, television broadcast, and multiperson dialogue. The existing problems include the following: the number of speakers is uncertain and there is no prior information about the number of speakers, the rotation of speakers is fast, and the length of each speaker's

speech is uncertain; there are a variety of noise in speech. How to solve these problems effectively and improve the robustness of the segmentation clustering system has become an important research direction, which is also the main research content of this paper.

In the work of speech recognition, the result of speech endpoint detection greatly affects the accuracy and rate of speech recognition segmentation [8]. Accurate endpoint detection can reduce a lot of computation for the feature extraction in the follow-up speech recognition and also make the acoustic model more accurate, so as to improve the accuracy of speech segmentation and recognition. Accurate endpoint detection of speech signal in complex background is a very important research branch in the field of speech recognition [9].

The so-called endpoint detection is to locate the speech segment in a section of original sound data and find the start- and endpoints of the speech segment [10, 11]. In order to eliminate the influence of channel and background noise, accurately determine the start- and endpoints of the sound segment, eliminate the silent segment in the speech signal, and make the energy of the whole speech signal concentrate on the sound segment, instead of being disturbed by background noise and silent segments, it can effectively improve the accuracy of speech segmentation and recognition. The performance, robustness, and processing time of a speech recognition system can be greatly improved by accurate and efficient endpoint detection. The traditional endpoint detection methods are mainly based on the characteristics of speech such as short-time energy, zero-crossing rate, etc. [12], but these characteristics are limited to the situation of no noise or high signal-to-noise ratio, and will lose the effect when the signal-to-noise ratio is low [13, 14].

According to the different ways of combination between segmentation and clustering, the current mainstream speech segmentation and clustering can be divided into two categories [15]: one is asynchronous strategy, that is, first segmentation and then clustering; in this strategy, segmentation and clustering are implemented step by step; the other is the synchronization strategy, that is, to complete speaker clustering while segmenting the speech of different speakers. ELISA proposed a typical speaker classification system in the literature [16], which combines two typical methods: one is based on asynchronous strategy, represented by the CLIPS system, which first automatically cuts audio into many small segments, so that each segment contains only one speaker, and then merges the same speakers through clustering; the other method is based on synchronous strategy, using the hidden Markov model (HMM) to achieve speaker clustering while segmentation. The LIA system is the representative of this kind of method. These two kinds of systems have their own advantages and disadvantages. The former is relatively simple, but the errors after each clustering may accumulate. The latter can correct the errors after each clustering, but it costs a lot of computing time, and cannot get enough training model.

Speech segmentation is an important part of asynchronous segmentation clustering, which includes speaker transform point detection and speech segmentation. The transform point detection is the key step of the segmentation module. The commonly used speaker speech segmentation methods are silence-based methods, metric-based methods, and model-based methods.

References [8, 15] proposed improved endpoint detection algorithms based on the combination of the energy and frequency band variance method and hybrid feature, respectively, in 2019. Reference [11] studied the speech endpoint detection method based on the fractal dimension method of adaptive threshold in 2020. In reference [17], cepstrum feature is used for endpoint detection, and cepstrum distance instead of short-time energy is used as threshold judgment, while speech detection based on the hidden Markov model is improved to adapt to noise changes. Reference [18] proposed a strong noise immunity VAD algorithm based on the wavelet analysis and neural network. The advantage of the algorithm based on silence is that the operation is relatively simple, and the effect is better when the background noise is not complex, but its limitations are exposed in the complex background, so some more effective algorithms have been proposed.

Document [19] studies the speaker transformation point detection with variable window length and realizes the online detection of transformation points, but its calculation is relatively large. Delacourt and Wellekens proposed a two-step speech segmentation algorithm, which first uses a fixed window to segment the speech initially and then merges the segmented speech segments. For different databases, this method has achieved good segmentation results. The advantage of speech segmentation based on distance scale is that it does not need any prior knowledge of speech, and the computational cost is low; the disadvantage is that it needs to set threshold according to experience, so the robustness and stability are poor, and it is easy to detect many redundant segmentation points.

The model method is to train the models of different speakers from the corpus and then use the trained model to classify the speech frame by frame, so as to detect the change points of speakers. Commonly used methods are as follows: universal background model (UBM) [20, 21], support vector machine (SVM) [22], and deep neural networks, (DNNs) [23]. The advantage of the model-based segmentation method is that it has higher accuracy than the distance-based method, but the disadvantage is that it requires prior knowledge, and the calculation cost is very high.

In the literature [24], the Gaussian mixture model is used in class modeling, which achieves high clustering purity. Document [25] studies the speaker clustering method based on the k-means algorithm, but the clustering results are greatly affected by the choice of initial cluster centers; if the choice is not appropriate, it may fall into local optimal solution, and the number of clusters K value needs to be given in advance.

To sum up, in order to improve the accuracy of speech endpoint detection, this paper proposes a new speech endpoint detection algorithm, which replaces the traditional double-threshold short-time average zero-crossing rate with a better spectral centroid feature, smoothes the feature curve

by median filter, and selects the threshold value by counting the local maxima of the feature sequence histogram. Compared with the traditional double-threshold algorithm, the proposed speech endpoint detection algorithm still has higher detection accuracy and noise immunity in low SNR.

The k-means algorithm has the advantages of convenient, fast calculation, and accurate results, but it needs to give the number of clusters in advance, and the results are greatly affected by the choice of the initial cluster center, so it is easy to fall into local optimum. The self-organizing neural network (SOM) has the advantages of strong explanatory, strong learning ability, and visualization, but the convergence speed is slow, and it cannot provide accurate clustering information, clustering accuracy for nonlarge volume of samples is poor. In order to seek better clustering means, the self-organizing neural network is introduced into speaker clustering, and an improved k-means speaker clustering algorithm based on self-organizing neural network is designed. The network is used to predict the number of clusters and the initial cluster centers of the k-means algorithm. The number of clusters is predicted by the winning situation of the neurons in the competitive layer of the trained network. The weight of neurons is used as the initial cluster centers of the k-means algorithm to cluster speakers. The experimental results of multispeaker mixed speech segmentation show that the improved clustering algorithm can not only make up for the shortcomings of the two algorithms but also improve the clustering accuracy.

2. Speaker Speech Segmentation Based on Improved Double-Threshold Endpoint Detection

2.1. Endpoint Detection Principle of Traditional Double-Threshold Method. The double-threshold endpoint detection method combines the short-time energy and the short-time average zero-crossing rate. Before the start of endpoint detection, two thresholds are set, respectively, for the short-time energy and the short-time average zero-crossing rate, and the thresholds are set empirically. The first is a low threshold, small value, more sensitive to signal changes, and more easily exceeded; the second is the high threshold, the value is large, and the threshold must reach a certain signal strength can be exceeded. Exceeding the low threshold does not mean the beginning of speech, which may be caused by short-term noise, and only exceeding the high threshold can basically determine the beginning of speech signal.

The whole speech signal can be divided into several segments: silence segment, transition segment, voice segment, and end segment. The basic steps of endpoint detection are as follows:

(1) In the silence segment, if one of the features of short-time energy or zero-crossing rate exceeds the low threshold, it will be marked as the beginning of the detection speech and enter the transition segment.

(2) In the transition stage, if the energy or zero-crossing rate characteristics of consecutive frames of speech exceed the high threshold, it is confirmed that they enter the real speech segment; otherwise, the current state is restored to the silent state.

(3) The endpoint of the speech segment can be detected reversely according to the above method. To sum up, the flowchart of double-threshold endpoint detection is shown in Figure 1.

2.2. Defects of Conventional Double-Threshold Method for Endpoint Detection. The ability to resist noise is weak. Noise environment is the main factor affecting the detection results, and different SNR and different noise will affect the accuracy of detection. Some noises contain rich high frequency components, and correspondingly the zero-crossing rate is relatively high. If the noise is too large, it will lead to a higher zero-crossing rate than vowels and initials in the noise of some silent segments. In the low SNR environment, the detection results are extremely unstable.

The threshold value is usually set by experience. It is extremely imprecise to use a fixed threshold to detect different speakers or different situations of speech.

Both the short-time energy and the short-time average zero-crossing rate are extracted in the time domain, so the calculation process is simple, and the actual characteristics of speech are not fully expressed.

The double-threshold method is generally used in speech recognition, which can only detect the beginning of a speech but cannot detect the internal pause of the speech.

Endpoint detection is used for speech segmentation, and the time domain of the corpus is larger than the short-time domain in speech recognition, so it is necessary to detect all the segmentation points in a long audio. Obviously, the traditional method cannot meet the requirements.

2.3. Improved Design of Double-Threshold Endpoint Detection Algorithm. In view of the defects of the traditional double-threshold method endpoint detection algorithm, the following three aspects are carried out to improve the detection method:

(i) In view of the limitation of the short-time average zero-crossing rate feature, the spectral centroid feature is used to replace it. The spectrum centroid is combined with short-time energy to detect

(ii) In order to improve the antinoise performance of the double-threshold method, the curves of the two features are smoothed by the median

(iii) In order to solve the problem of poor accuracy caused by the threshold selection based on experience, an algorithm is proposed to select the threshold reasonably by analyzing the whole feature sequence

2.3.1. Spectral Centroid Characteristics. Spectral centroid is a parameter describing the property of timbre. Different from short-time energy and short-time average zero-

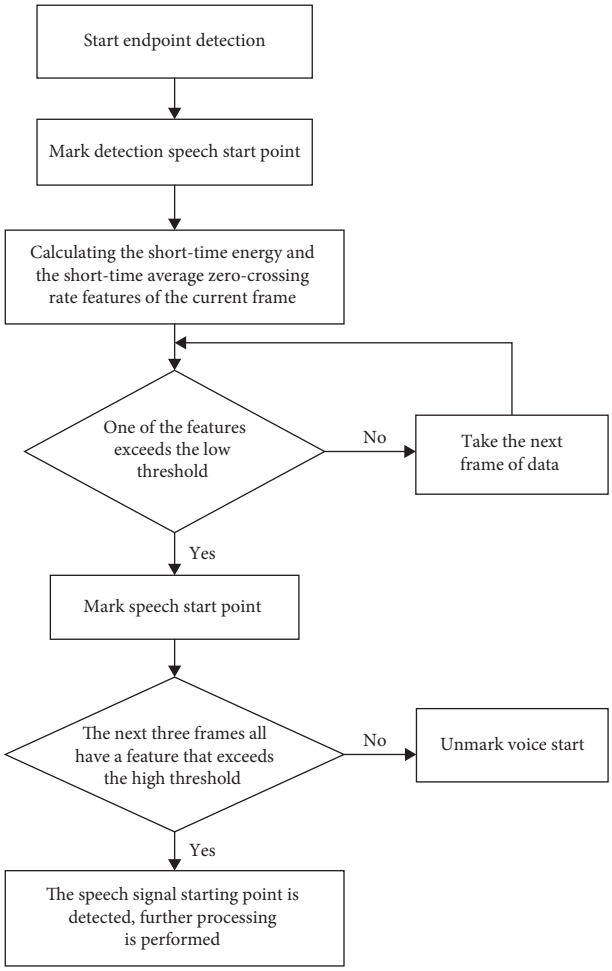

Figure 1: Traditional flowchart of the double-threshold method VAD.

crossing rate, spectrum centroid is a characteristic parameter extracted in frequency domain. First, short-time Fourier transform must be done to the signal, and then time-frequency analysis must be done. After getting the spectrogram of the signal, the spectral centroid C_i of the speech of the i-th frame is

$$C_i = \frac{\sum_{k=1}^{N} (k+1) X_i(k)}{\sum_{k=1}^{N} X_i(k)}. \quad (1)$$

In the formula, $X_i(k)$ is the k-th discrete Fourier transform (DFT) coefficient in the spectrogram of the speech of the i-th frame, and the visible spectral centroid represents the center of spectral gravity and is the concentration point of spectral energy, generally speaking, the smaller the spectral centroid, the more concentrated the energy is in the low frequency range.

The main reasons to select the combination of short-term energy and spectral centroid features for endpoint detection are as follows: for simple cases (background noise is not very high), the short-term energy of speech segment is usually greater than that of nonspeech segment. Spectral centroid is a feature in the frequency domain, which can reflect the frequency information of the signal more accurately than the short-time average zero-crossing rate in the time domain. If the nonspeech segment includes simple ambient noise, then the spectral centroid of the noise is usually lower than that of the speech segment.

2.3.2. Median Filter Smoothing Method. After extracting the features of short-time energy and spectral centroid, it is defective to set the threshold value on the feature curve directly when detecting speech, because when the signal-to-noise ratio is low, the fluctuation of the feature curve in nonspeech segments is large, and the low threshold value will easily lead to misjudgment, while the high threshold value will lead to undetected. Therefore, it is necessary to reduce the fluctuation of the feature curve in the nonspeech segment, and median filtering can be used to smooth the curve.

Median filtering is a nonlinear smoothing technique based on statistical ordering theory. The basic idea is to find the closest element to its surroundings for any signal element (sound or image). The principle is to replace the value of a point in the signal sequence with the median value of each point in its neighborhood, so as to eliminate the isolated noise point.

2.3.3. Threshold Selection Algorithm. After the median smoothing filtering, the short-time energy and spectral centroid characteristic curves are smoothed. The traditional double-threshold method is to set the threshold by experience, but the speech characteristics of different people or different situations are very different, using the same threshold to filter speech is very inaccurate.

Therefore, a new algorithm is designed, which can select the threshold dynamically and reasonably to improve the detection accuracy in the case of noise.

First, the histogram of the smoothed feature sequence is calculated. The histogram is an accurate graphical representation of the distribution of data and the estimation of the probability distribution of variables. In order to establish the histogram, the first step is to segment the range of values, usually at equal intervals, and then count the number of times the data appear in each segment.

Taking the spectral centroid characteristic sequence as an example, the minimum and maximum values of spectral centroid characteristic coefficients are first found out, the range from the minimum to the maximum is divided into L sections averagely, the frequency of spectral centroid coefficients appearing in each section is counted, and finally the histogram is drawn. Let the value of item I ($i = 1, 2, \ldots, L$) in the histogram be $f(i)$.

The local maximum value M of the statistical histogram is due to the fact that in a certain position, and if the probability of the occurrence of the characteristic sequence is much greater than that of the adjacent position, then it is very likely that the place is the transition from nonspeech to speech. The basic principle is as follows: if a segment appears more times than the adjacent segments in the histogram, the characteristic coefficient value corresponding to the center

of the segment is a local maximum. Figures 2 and 3 show histograms of short-time energy and spectral centroid signature sequences, respectively.

The specific statistical methods are as follows:

Set a step length step, and judge from the first item in the histogram to the (L-step) item in turn, when $i \leq$ step, if it appears

$$\text{mean}(f(1{:}\ i)) < f(i)\ \&\&\ \text{mean}(f(i+1{:}\ i+\text{step})) < f(i). \tag{2}$$

Then, the characteristic coefficient corresponding to the center of the i-th segment in the histogram is a local maximum. When $i >$ step, if appears

$$\text{mean}(f(i-\text{step}{:}\ i-1)) < f(i)\ \&\&\ \text{mean} \cdot (f(i+1{:}\ i+\text{step})) < f(i). \tag{3}$$

Then, the characteristic coefficient corresponding to the center of the i-th segment in the histogram is a local maximum. According to the above statistical method, let the number of detected maximum values be n and the threshold value of the characteristic sequence be T. The calculation of T is divided into the following three cases:

(1) $n = 0$; then,

$$T = \frac{\sum_{k=1}^{N} C_k}{4N}, \tag{4}$$

where C_k is the k-th value of the feature sequence, this expression means that if the local maximum is not detected from the beginning to the end, the threshold value is replaced by (1/4) of the average value of the feature sequence, but this case is not common.

(2) $n = 1$; then,

$$T = M, \tag{5}$$

where M is the only detected local maximum, and this is not often the case. Usually, more than two local maxima are detected.

(3) $n = 2$; then,

$$T = \frac{W \cdot M_1 + M_2}{W + 1}. \tag{6}$$

Arrange all the detected maximum values in descending order of frequency. In equation (3) above, M_1 and M_2 are the first two maximum values. W is a user-defined parameter, and the higher the W, the closer the threshold value is to the first maximum value M_1.

The thresholds of short-time energy and spectral centroid characteristics, denoted as T_1 and T_2, respectively, are calculated by this method. When two features in a frame of the audio signal are higher than the threshold value, it is judged that the frame is a speech signal.

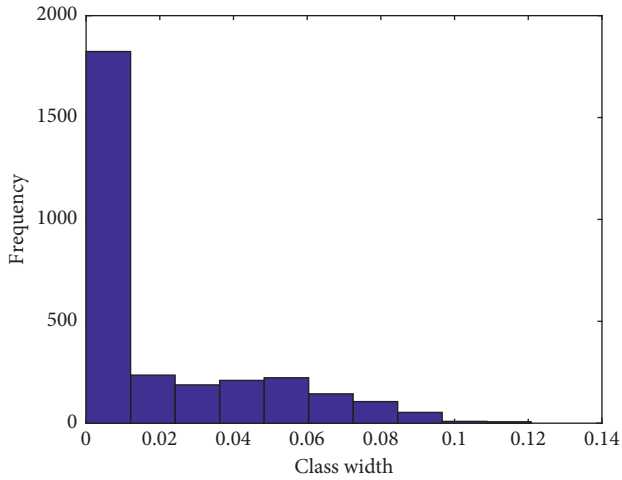

Figure 2: Histogram of characteristic series of short-time energy.

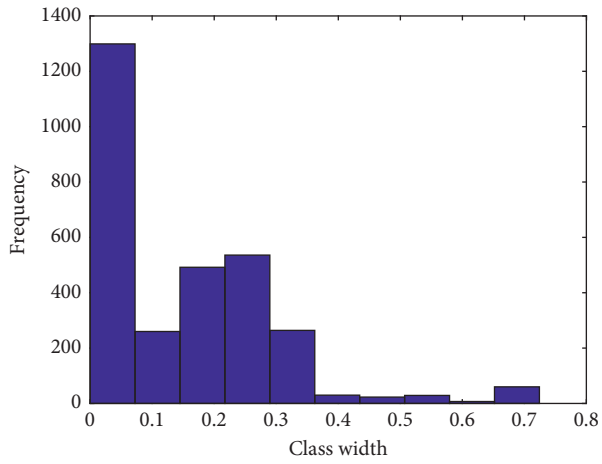

Figure 3: Spectral centroid characteristic sequence histogram.

2.3.4. Speaker Speech Segmentation Based on Improved Double-Threshold Method. The improved detection process is as follows:

(i) The speech signal is collected, and the time domain waveform is obtained.

(ii) The speech is divided into frames and windows, and the short-time Fourier transform is performed to obtain the spectrogram of the signal.

(iii) The short-time energy feature E_n is extracted in the time domain and the spectral centroid feature C_n are extracted in the frequency domain.

(iv) The short-time energy feature and the spectrum centroid feature are smoothed by median filtering twice.

(v) The histograms of the above two feature sequences are calculated, respectively, and the local maxima of the histograms are counted, and the threshold values of the two features are calculated. The threshold value of short-time energy feature is T_1, and that of spectral centroid feature is T_2.

(vi) If the short-time energy feature of a frame is greater than T_1 and the spectrum centroid feature of the frame is greater than T_2, the frame is marked as a speech frame; otherwise, it is marked as a nonspeech frame.

(vii) Postprocessing stage (use according to the situation): extend the two ends of each voice segment by 2 windows, and finally merge the continuous segments as the final voice segment.

The speaker speech segmentation algorithm based on the improved double-threshold method is shown in Figure 4.

Among them, the postprocessing stage is mainly to take into account the extremely short pauses that sometimes occur in speech, eliminating these pauses and merging the speech can reduce the voice segments and reduce the complexity of the results. However, in a few cases, these short pauses may also be the change point of the speaker, which will lead to wrong merging and affect the next stage of speech clustering. Therefore, the postprocessing method is used when the audio contains only one person's voice, but not when there is a multiperson conversation.

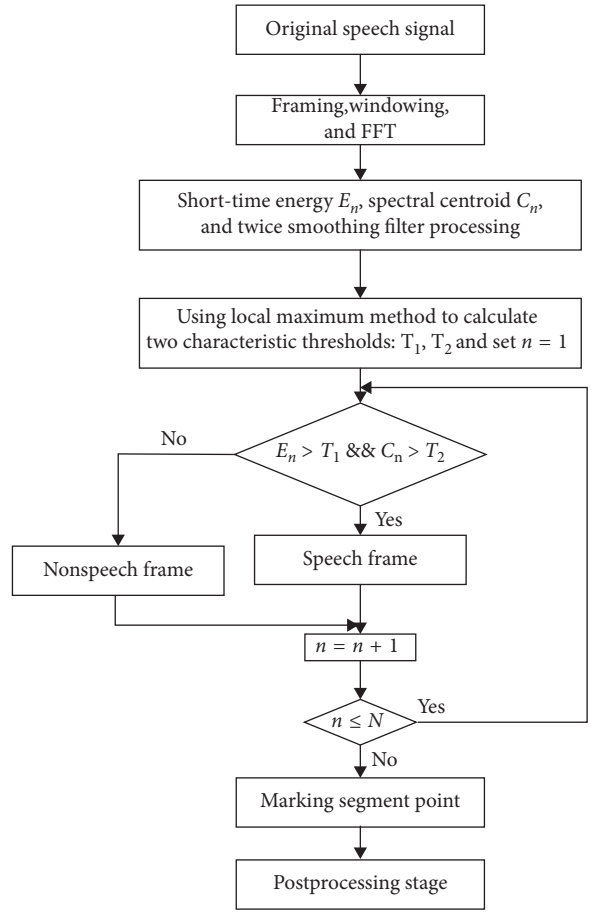

Figure 4: Speaker segmentation algorithm flowchart based on improved double-threshold method VAD.

2.4. Comparative Experimental Analysis. The experiment of endpoint detection of speech signal is carried out by using Matlab software, and the data are recorded by the Newsmy recorder. The experiment sample is a 1.5 s speech, and the content is the Chinese pronunciation of Ni Hao. The output is a standard Windows WAV audio file, and the file name is Hello. wav, sampling frequency is FS = 8 kHz and monophonic, using 16 bit encoding. For the original speech, we use the traditional double-threshold method and the improved method to carry out endpoint detection experiments, and make a comparative analysis.

Extract the time domain waveform of the Hello. wav raw audio file as shown in Figure 5.

Firstly, the speech signal is divided into frames and windowed. The frame length is wlen = 200 (each frame has 200 sampling points), the frame shift is inc = 100, and the window function is Hanning window. At the sampling rate of fs = 8 khz, the total number of sampling points of the speech sequence is 12001, which is divided into 119 frames, and the corresponding time of each frame is 25 ms. Calculate the energy of each frame of speech, and extract the short-time energy characteristics of speech. Figure 6 shows the short-time energy map of speech.

The short-time average zero-crossing rate of each frame is calculated, and the zero-crossing rate feature is extracted. Figure 7 shows the short-time average zero-crossing rate characteristics of speech.

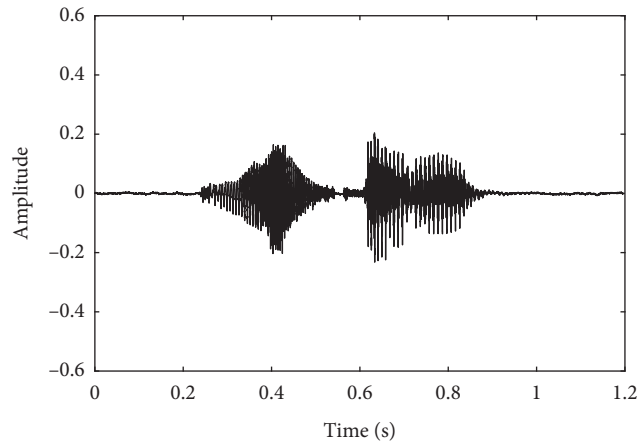

Figure 5: Original speech signal.

2.4.1. Analysis of Endpoint Detection Based on Traditional Double-Threshold Method. Combining the short-time energy with the short-time average zero-crossing rate, based on

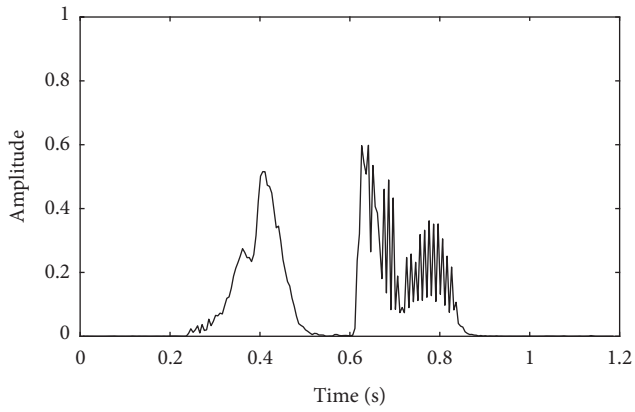

Figure 6: Short-time energy of original speech signal.

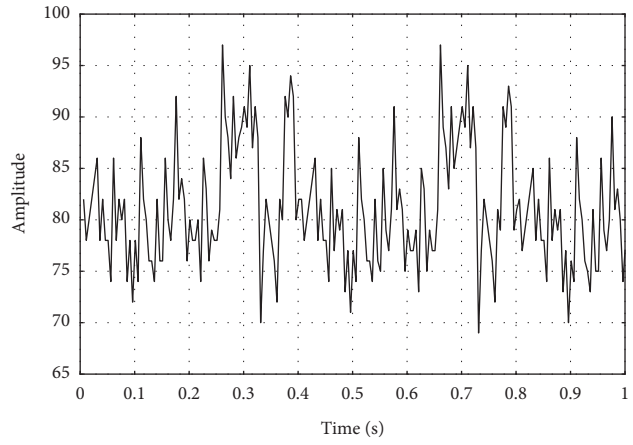

Figure 7: Short-time average zero-over rate of original speech signal.

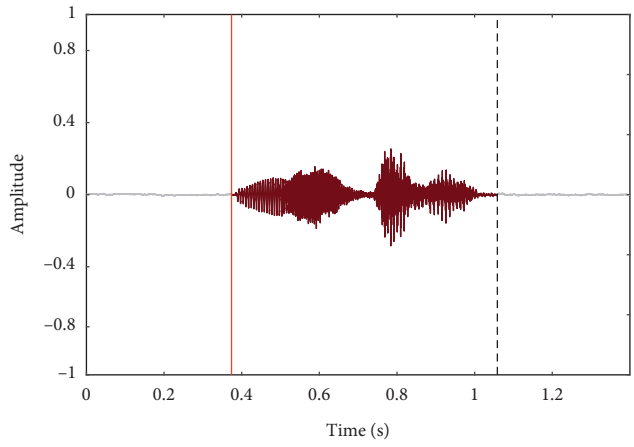

Figure 8: Results of double-threshold method VAD.

the traditional double-threshold endpoint detection algorithm, the location of hello speech in the time domain waveform is detected, and the detection result is shown in Figure 8.

In Figure 8, the beginning of the speech is marked with a solid line and the end is marked with a dashed line. It can be seen from the picture that the speech starts at about 0.38 s and ends at about 1.03 s, which is consistent with the actual situation. The results show that the original double-threshold endpoint detection method can achieve good detection results in extremely low noise environment.

2.4.2. Analysis of Endpoint Detection Based on Improved Double-Threshold Method. First, the spectral centroid of each frame is calculated, and the spectral centroid feature is extracted; then, the short-time energy and spectral centroids are smoothed by median filtering twice, and the threshold values of the two features are calculated simultaneously. The endpoint detection results of the improved double-threshold method are shown in Figure 9:

Figures 9(a) and 9(b) show short-time energy and spectral centroid feature images, respectively, and the solid line part is the original feature curve, and the dashed line part is the feature curve after two times of smooth filtering. The ordinate corresponding to the black thick bar in the figure is the characteristic threshold value selected after calculation. If the feature curve exceeds the bar, the feature exceeds the threshold, and only when both features exceed the threshold can the frame be judged as a voice frame.

Figure 9(c) shows the endpoint detection result of the improved algorithm, wherein the beginning of the speech is marked with a solid line and the end is marked with a dashed line. It can be seen from the picture that the method detects two segments of speech, which appear at 0.74–1.06 s and 1.08–1.26 s. In practice, there is a slight pause between the word you and the word good in the audio. The overall appearance time of speech detected by the two methods is basically the same, but the original double-threshold method only detects the overall speech segment, while the improved method can accurately detect the pause in the middle of the speech segment. Therefore, the improved method can better meet the needs of speaker segmentation in this paper.

2.4.3. Comparison of Detection Accuracy between Two Methods. Aiming at the original audio file of Hello. wav and the audio file with different degrees of Gaussian white noise, the original double-threshold method and the improved double-threshold method are used to detect endpoints. The formula for endpoint detection accuracy is as follows:

$$\text{Accuracy} = \frac{\text{total frame number} - \text{number of error frames}}{\text{total frame number}}. \tag{7}$$

For the speech with different noise levels, the endpoint detection accuracy is calculated, and the accuracy results are shown in Figure 10.

From the accuracy of Figure 10, we can see that both detection algorithms can accurately detect the endpoints of speech in the case of silence or very small noise. When different levels of noise are imposed on the audio files, with

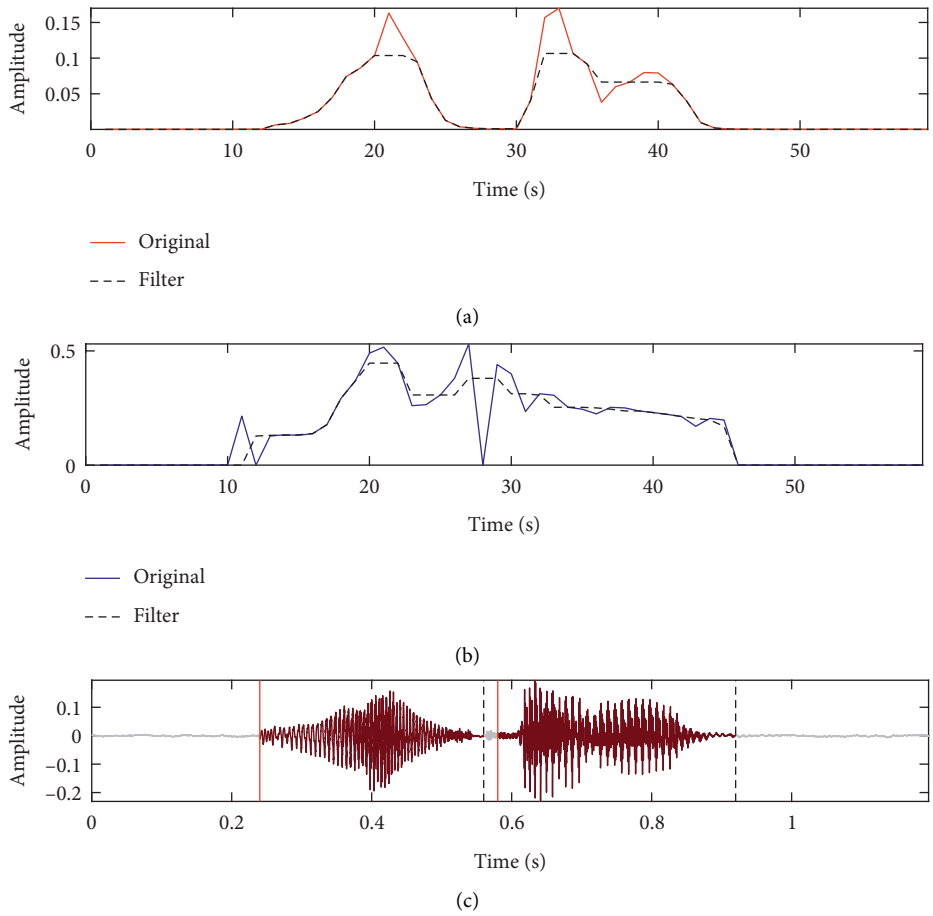

Figure 9: Results of the improved double-threshold method VAD: (a) short-time energy, (b) spectral centroid, and (c) results of improved double-threshold method VAD.

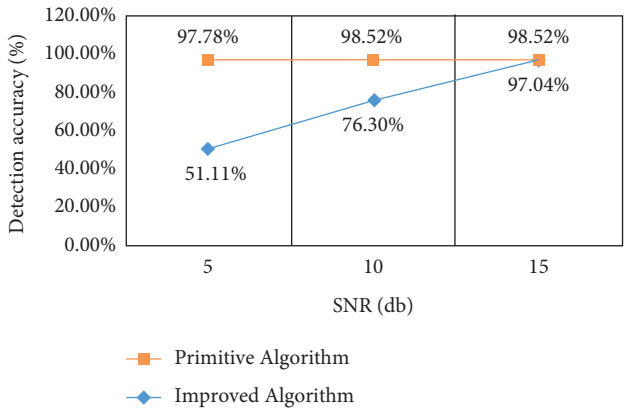

Figure 10: Accuracy of endpoint detection based on two algorithms in Gaussian white noise.

the continuous reduction of the SNR, the noise is continuously enhanced, and the detection accuracy of the traditional double-threshold method is significantly reduced, while the improved algorithm can still maintain a high detection accuracy.

3. Improved *K*-Means Speaker Speech Clustering Based on Self-Organizing Neural Network

Clustering technology belongs to the typical unsupervised learning, that is, the given data only have features without labels, and it is classified by the internal relationship and similarity between the data. On the contrary, supervised learning means that has given the training data contains labels and features, and we can find the relationship between features and labels through training, so that we can judge the label when facing new data. A comparison of the system components of the two learning methods is shown in Figure 11.

Figures 11(a) and 11(b) show system compositions of a supervised learning mode and an unsupervised learning mode. It can be seen that as an unsupervised learning method, clustering does not need to set the output in advance, there is no human interference, and its purpose is to bring similar objects together, regardless of what this class is. In this paper, the clustering technology is applied to the classification of the speaker's speech, and the speech of the same person is classified into one class by clustering.

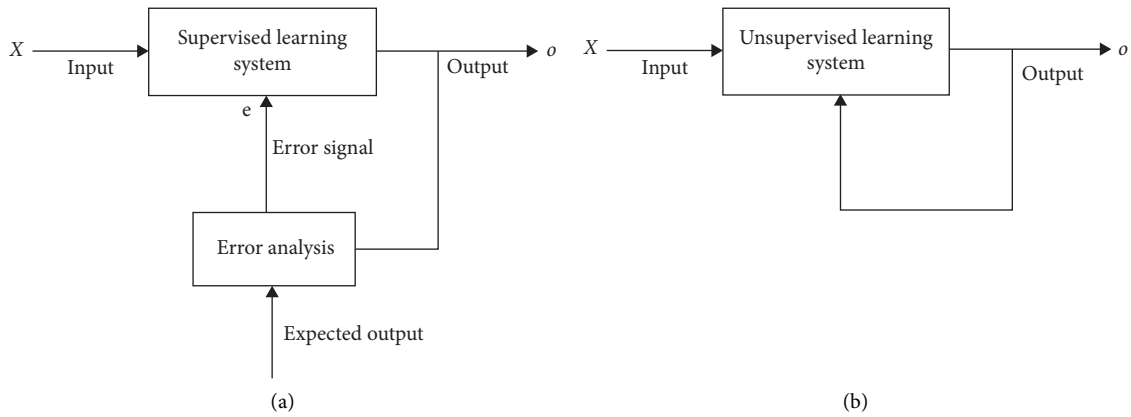

Figure 11: Supervised learning and unsupervised learning.

The k-means algorithm and self-organizing neural network (SOM) algorithm are widely used in clustering analysis. The k-means algorithm has the advantages of convenient, fast calculation, and accurate results, but it needs to give the number of clusters in advance, and the results are greatly affected by the choice of initial cluster centers, so it is easy to fall into local optimum. The self-organizing neural network has the advantages of strong explanatory, strong learning ability, and visualization, but its convergence speed is slow, and it cannot provide accurate clustering information, and the clustering accuracy is poor in nonlarge volume of samples, so it is not suitable for speaker speech clustering.

Therefore, in order to seek better clustering means, this paper introduces self-organizing neural network into speaker clustering and uses it to improve the k-means algorithm, through the network to predict the number of k-means algorithm clustering and the initial cluster center, in order to overcome the shortcomings of these two methods and improve the clustering accuracy.

3.1. Self-Organization Neural Network. The self-organizing feature map (SOM) neural network is based on the phenomenon of lateral inhibition in the biological neural system. The basic idea is that for a specific input pattern, each neuron competes for the opportunity to respond, and ultimately only one neuron wins, and the winning neuron represents the classification of the input pattern. Therefore, the self-organizing neural network can be easily associated with clustering.

The structure of self-organizing neural network is generally a two-layer network: input layer + competition layer, and there is no hidden layer, sometimes there are lateral connections between neurons in the competitive layer. A typical self-organizing neural network structure is shown in Figure 12.

Input layer: simulate the retina which perceives external information, receives information, plays the role of observation, and transmits the input mode to the competition layer. The number of neurons in the input layer is generally the number of samples.

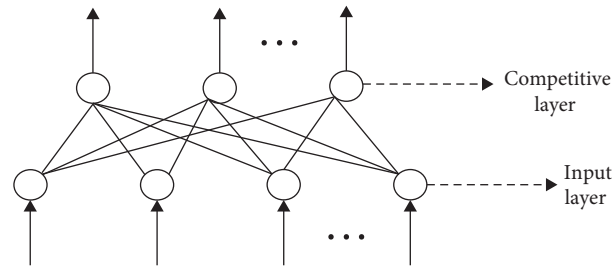

Figure 12: Typical SOM network model.

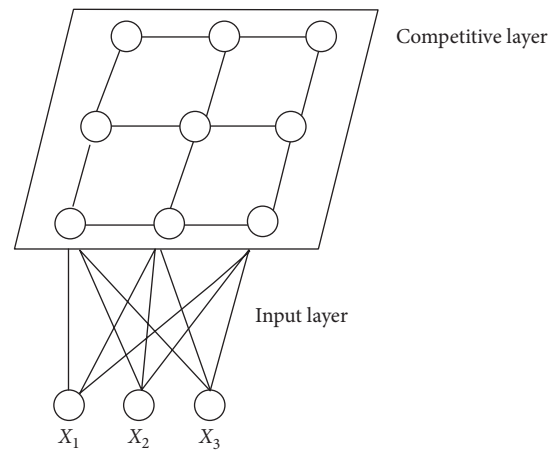

Figure 13: Two-dimensional SOM network model.

Competition layer: simulate the responding cerebral cortex, which is responsible for comparative analysis of input, looking for rules and classification. The output of competition layer represents the classification of the pattern, and the number of neurons is usually the number of categories.

Another structure is the two-dimensional form, which has a more cortical image, as shown in Figure 13.

Each neuron in the competition layer links laterally with its nearby neurons in a certain way, forming a plane, similar to a checkerboard. In this kind of structure, the neurons in the competition layer are arranged as a two-dimensional node matrix, and the neurons in the input layer and output layer are connected with each other according to the weights.

3.2. Competitive Learning Rule. The self-organizing neural network follows the rule of competitive learning, that is, competitive winning neurons will inhibit the losing neurons. Because it belongs to unsupervised learning, there is no output of the desired pattern in the sample. There is no a priori knowledge about which class an input element should be classified into, so it is necessary to classify according to the similarity between samples, which is the basis of self-organizing neural network clustering.

The basic steps of competitive learning rules are as follows:

(1) Vector normalization

The input vector X of the self-organizing neural network and the weights $W_j(j = 1, 2, ..., m)$ of each neuron in the competition layer are all normalized to obtain $\widehat{X}$ and $\widehat{W}_j$:

$$\widehat{X} = \frac{X}{\|X\|}, \qquad \widehat{W}_j = \frac{W_j}{\|W_j\|}. \tag{8}$$

(2) Find the winning neuron

Comparing $\widehat{X}$ with the weights W_j of all neurons in the competition layer, the most similar neuron is the winning neuron, and its weight is $\widehat{W}_{j^*}$.

$$\left\|\widehat{X} - \widehat{W}_{j^*}\right\| = \min_{j\in\{1,2,\ldots,m\}}\left\{\left\|\widehat{X} - \widehat{W}_j\right\|\right\}. \tag{9}$$

As said before, the normalized similarity is the largest; that is, the inner product is the largest:

$$\widehat{W}_{j^*}^T\widehat{X} = \max_{j\in\{1,2,\ldots,m\}}\left(\widehat{W}_j^T\widehat{X}\right). \tag{10}$$

It is equivalent to finding the point with the smallest angle in the unit circle.

(3) Network weight adjustment

According to the learning rule, the output of the winning neuron is 1, and the output of the other neurons is 0; that is,

$$y_j(t+1) = \begin{cases} 1, & j = j^*, \\ 0, & j \neq j^*. \end{cases} \tag{11}$$

Only the winning neuron has the right to adjust the weight vector as follows:

$$\begin{cases} W_{j^*}(t+1) = \widehat{W}_{j^*}(t) + \Delta W_{j^*} = \widehat{W}_{j^*}(t) + \eta(t)\left(\widehat{X} - \widehat{W}_{j^*}\right), & j = j^*, \\ W_j(t+1) = \widehat{W}_j(t), & j \neq j^*. \end{cases} \tag{12}$$

Among them, $0 < \eta(t) < 1$ is the learning rate, which generally decreases with time; that is, the degree of adjustment becomes smaller and smaller, and gradually tends to cluster centers renormalization.

After adjustment, the weight vector is no longer a unit vector, so it needs to be normalized again, and the network needs to be retrained until the learning rate $\eta(t)$ attenuates to zero, and the algorithm ends

In the testing phase, the inner product of the given object and the weights of each neuron is calculated, and the most similar neuron is assigned to which class.

The Kohonen algorithm is usually used for the two-dimensional self-organizing neural network structure. This algorithm is an improvement of the above competitive learning rules. The main difference between Kohonen algorithm and competitive learning rule is that the way of lateral inhibition of neuron weight adjustment is different. In the competitive learning rule, only the winning neuron has the right to adjust the weight. In Kohonen algorithm, the influence of the winning neuron on the surrounding neurons is from near to far, from excitement to inhibition, so the nearby neurons also need to adjust their weights to varying degrees under its influence. Take the winning neuron as the center, set a neighborhood radius R, and this range is called the dominant neighborhood. In the algorithm, the neurons in the winning neighborhood adjust their weights according to the distance from the winning neuron. At the beginning, the radius of the winning neighborhood is set to be very large, and as the number of training increases, the size shrinks until it is zero, as shown in Figure 14.

3.3. Design of Improved k-Means Speaker Clustering Algorithm Based on Self-Organizing Neural Network. The operation of the self-organizing neural network is divided into two stages: training and testing. In the training stage, the input training set samples, for a specific input, the competition layer will have a neuron to produce the largest response to win. The neural network adjusts the weights by training samples in a self-organizing way and finally makes some neurons in the

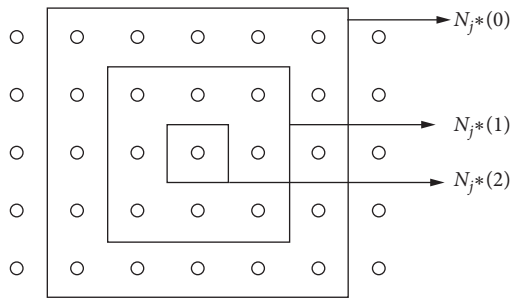

Figure 14: Contraction of superior neighborhood.

competition layer sensitive to the input of a specific pattern class, and the corresponding weights become the center of each input pattern. Thus, the characteristic graph of the distribution of the reactive class is formed in the competitive layer.

The k-means method has two shortcomings: the number of clusters needs to be given in advance and the selection of initial clustering centers is very dependent on the algorithm. The self-organizing neural network has the advantages of strong learning ability, strong interpretation, visualization, and so on. However, it also has the limitations of long training time, slow convergence rate, and unsatisfactory clustering results for small data.

In this paper, the self-organizing neural network is introduced into the k-means algorithm. The improved algorithm can not only make up for the slow convergence of self-organizing neural network but also improve the k-means algorithm:

(1) Predictive clustering number

Firstly, a self-organizing neural network is used to train the speech feature set for a short period of time, and a discrimination method is designed to determine the number of classes K according to the winning situation of the neurons in the competitive layer of the network.

(2) Finding initial clustering center

The weight of the neuron is used as the initial clustering center, and the k-means algorithm is used to complete the speech segment clustering. In the improved algorithm, self-organizing neural network is used to get the initial value of the k-means algorithm, which makes it unnecessary to wait for the complete convergence of the network and reduces the training time of the network. For the trained network, the more times the neurons in the competition layer win, the closer to the actual clustering center. Therefore, the number K of clusters can be predicted and the initial cluster centers can be calculated by the winning situation of neurons.

The specific steps of the algorithm are as follows:

(1) Sample input

Based on the improved double threshold endpoint detection algorithm, the long audio is segmented into n short time speech segments containing only one person's speech, thus extracting the MFCC features of each speech segment to form a feature set $X_i (= 1, 2, \ldots, n)$ as the input of the system

(2) Training self-organizing neural network

(a) Let us start with a rough estimate of the number of categories K, assuming no more than nine speakers in this experiment $k \le 9$. Nine neurons are set in the competition layer of the network, and the 3×3 layout is adopted. The number of neurons in the input layer is n.

(b) Initialization: the speech segment feature vector is normalized to obtain $\widehat{X}_i (i = 1, 2, \ldots, n)$, and the neuron weight $W_j (j = 1, 2, \ldots, 9)$ in the competition layer is assigned a smaller random number, and normalized to obtain hat $\widehat{W}_j (j = 1, 2, \ldots, 9)$. The initial values of the dominant neighborhood $\{N_{j*}\}(0)$ and the learning rate $\backslash\eta$ are set. Let the training time $t = 1$ because the self-organizing neural network is used in the front end of the k-means method in this algorithm, reduce the training time, we do not need to wait for the network to converge completely. We only need to set a relatively small number of iterations (100 times in the experiment).

(c) Finding the winning neuron: for the i-th input object, calculate the inner product of $\widehat{X}_i$ and $\widehat{W}_j$ find out the neuron corresponding to the maximum inner product, which is the winning neuron j^*:

$$\widehat{W}_{j^*}^T \widehat{X}_i = \max_{j \in \{1,2,\ldots,m\}} \left(\widehat{W}_j^T \widehat{X}_i \right), \tag{13}$$

where W_j is the neuron weight.

(d) Defining the dominant neighborhood $N_{j^*}(t)$ taking j^* as the center and determining the dominant neighborhood $N_{j^*}(t)$ at time t, generally the initial neighborhood $N_j^*(0)$ is larger (about 50%–80% of the total nodes), and $N_{j^*}(t)$ decreases with the increase of training time.

(e) Adjusting the weights: adjusting the weights of all neurons in the superior neighborhood N:

$$W_j(t+1) = W_j(t) + \eta(t, N)\left[x_i - W_j(t)\right], \quad i = 1, 2, \ldots, n,\ j \in N_{j^*}(t), \tag{14}$$

where the learning rate $\eta(t, N)$ is a function of the training time t and the topological distance N between the winning neuron j^* and superior neighborhood neuron j.This function generally has the following rules:

$$t\uparrow \to \eta\downarrow, \quad N\uparrow \to \eta\downarrow. \tag{15}$$

Example:

$$\eta(t, N) = \eta(t)e^{-N}, \tag{16}$$

where $\eta(t)$ can take the monotone decreasing function of t, also called annealing function.

(f) $t = t + 1$, steps (c) to (e) are repeated until $\eta(t) \le \eta_{\min}$ or the maximum number of training times is reached, and step (3) is entered.

(3) K value decision

The winning times of each neuron in the competition layer after training were as follows: P_j ($j = 1, 2, \ldots, 9$). Let $k = 0$ and $j = 1$, if

$$P_j > \frac{4}{3}\text{mean}[P_1, P_2, \ldots, P_9]. \tag{17}$$

Then, the number of categories $k = k + 1$ and $j = j + 1$.

Continue to judge according to formula (17), and the final number of categories is obtained as follows: $k = k_0$.The idea here is that the more times a neuron wins, the closer it is to the actual clustering center. Neurons with fewer wins (less than the average number of wins) are considered to be far away from the cluster center and ignored.

(4) Initial cluster center prediction

Retraining the self-organizing neural network: at this time, k_0 neurons are set in the competition layer, and other things remain unchanged. When the network training is finished, the weight value W_l($l = 1, 2, \ldots, k_0$) of each neuron is obtained, which is used as the initial clustering center in the k-means method.

(5) K-means speaker cluster

(a) The input of the algorithm is as follows: MFCC feature set X_i ($i = 1, 2, \ldots, n$) of speech segment, class number K_0, and initial clustering center μ_j:

$$\mu_j = W_j, \quad (j = 1, 2, \ldots, k_o). \tag{18}$$

(b) Class partition C is initialized to

$$C_j = \varphi, \quad j = 1, 2, \ldots, k_o. \tag{19}$$

(c) The distance between each sample X_i and each cluster center μ_j is calculated as

$$d_{ij} = \left\| X_i - \mu_j \right\|_2^2. \tag{20}$$

For X_i, it is assigned to the class λ_i corresponding to the smallest d_{ij}, and the class division is updated:

$$C_{\lambda_i} = C_{\lambda_i} \cup \{X_i\}. \tag{21}$$

(d) For $j = 1, 2, \ldots, k_o$, recalculate the cluster centers for all sample points in C_j:

$$\mu_j = \frac{1}{N_j}\sum_{i=1}^{N_j} X_i, \tag{22}$$

where N_j is the number of samples in each category C_j.

(e) $t = t + 1$, and the error square sum criterion function is used to explicitly determine whether the algorithm is finished or not:

$$J = \sum_{j=1}^{k_o}\sum_{x \in C_j}\left(x - \mu_j\right)^2. \tag{23}$$

If $|J(n) - J(n-1)| < \xi$ is satisfied, or the number of iterations $t = T$, the algorithm ends and goes to step (6). Otherwise, go to step (b).

(6) Algorithm output

Output cluster partition $C = \{C_1, C_2, \ldots, C_{k_o}\}$, and the algorithm ends.

To sum up, the flow of the improved k-means speaker clustering algorithm based on the self-organizing neural network is shown in Figure 15.

3.4. Experimental Analysis. The experiment sample is several minute long multiperson dialogue audio, uses Newsmy recording pen to record, simulates the multiperson meeting the situation. The output is a standard Windows WAV audio file, and sampling frequency is fs = 8 kHz, monophonic, using 16 bit encoding. Recording audio requires crosstalk, and in order to ensure the purity of voice and improve clustering accuracy, speak clearly and do not send out cough and other noise.

The experiment process is shown in Figure 16. The k-means speaker clustering algorithm, the self-organizing neural network speaker clustering algorithm, and the improved k-means speaker clustering algorithm based on the self-organizing neural network are used to cluster the speech segments, and the effectiveness of the improved algorithm is verified by comparative analysis.

Select an audio file named Recording 1.wav, which lasts for 3 minutes and contains the voices of two men and one woman.

Extract the time domain waveform of the Recording 1. wav audio file as shown in Figure 17.

Firstly, the speech signal is preprocessed, including preemphasis, and subframe and window processing. Frame length wlen = 200, frame shift inc = 100, and window function is Hanning window. The duration of the audio sequence is 180 s, and the total number of sampling points is 1419856 at the sampling rate of FS = 8 khz. The sequence is divided into 14197 frames, and the corresponding time of each frame is 25 ms. Through the time domain waveform, we can see that the audio has a number of voice segments, and there is a short gap between the voice segments.

The short-time energy and spectral centroid characteristics of each frame of speech are calculated from beginning to end. The audio is segmented based on the improved

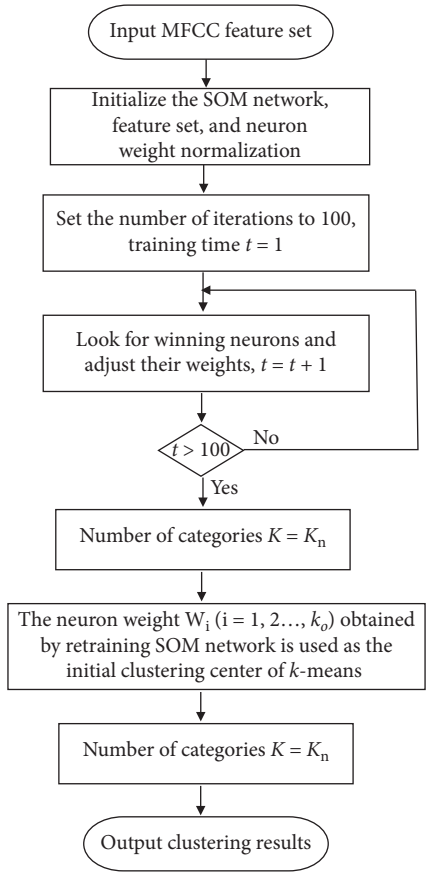

Figure 15: Flowchart of the SOM + k-means speaker clustering algorithm.

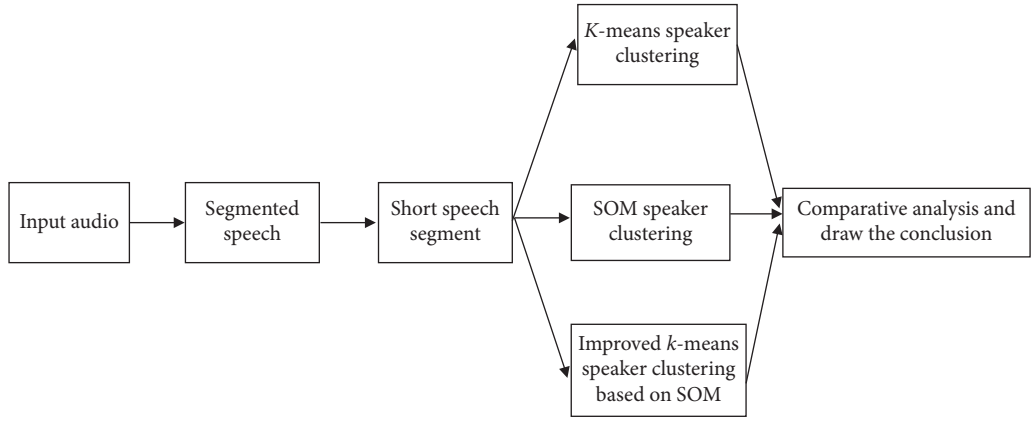

Figure 16: Flowchart of speaker clustering experiment.

double-threshold endpoint detection method, and the segmented speech waveform is shown in Figure 18.

As can be seen from Figure 18, the audio is divided into a number of speech segments. In the picture, the speech segments are shown in dark colors, and the silence between the speech segments is shown in light gray. After speech segmentation, a total of 96 short-time speech segments are obtained, and each of which contains only one person's speech.

In the clustering experiment, MFCC (mel-frequency cepstrum coefficient) is used as the basis to distinguish different speakers.

The average of MFCC vectors of all frames in the speech is used to represent the MFCC feature of the whole speech; that is, the MFCC feature vector is obtained by calculating the average of the feature matrix according to the column.

For these 96 speech segments, the MFCC feature vectors of each speech segment are extracted, respectively, and the

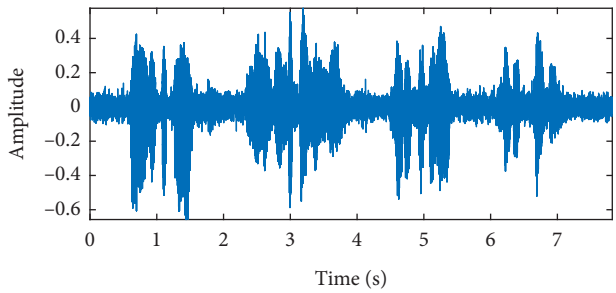

Figure 17: Partial audio waveform.

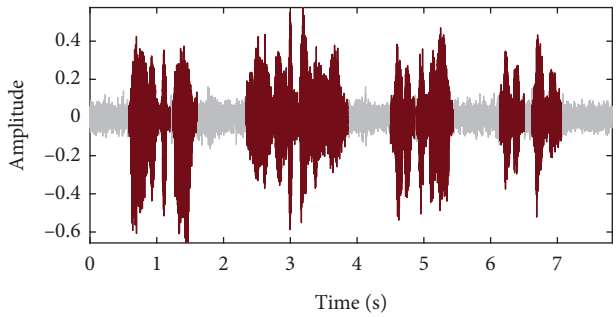

Figure 18: Segmented speech (results of the improved double-threshold method VAD).

feature set is synthesized. In data processing, different evaluation indexes usually have different dimensional units, which affects the analysis results. In order to eliminate the influence of different dimensions and make them comparable, it is necessary to normalize the feature set. The normalized data are between −1 and 1, which are in the same order of magnitude, and are suitable for comprehensive evaluation. The normalized MFCC feature set is used as the input sample of the clustering system.

The MFCC feature set of the audio is shown in Figure 19.

The feature set has 12 columns, and the dimension of MFCC is 12. Each line represents a sample, that is, a total of 96 short-duration speech segments.

3.4.1. K-Means Speaker Cluster. For feature set $X_i (i = 1, 2, \ldots, 96)$, k-means speaker clustering algorithm, self-organizing neural network speaker clustering algorithm, and improved k-means speaker clustering algorithm based on self-organizing neural network are used for clustering experiments.

First, listen to the segmented audio and attach distinguishing labels for different speakers to facilitate subsequent comparative analysis. Among them, Zhang San's pronunciation is expressed by "a," Li Si's pronunciation is expressed by "b," and Wang Wu's pronunciation is indicated by "c." Table 1 shows the speech category.

Based on the k-means speaker clustering algorithm, the MFCC feature set is clustered X_i $(i = 1, 2, \ldots, 96)$, K value is set to 3, and the cluster center is initialized randomly. When the initial clustering centers are selected properly, the highest clustering accuracy can reach 94.8%. However, after 50 times of k-means clustering for this sample, there are 12 times of abnormal clustering due to improper selection of initial clustering centers, which greatly reduces the average clustering accuracy. The average clustering accuracy of these 50 k-means was 84.5%. Table 2 describes the abnormal clustering results when the initial clustering center is not selected properly. The suffix "×" is the wrong clustering item, and the accuracy of this clustering is 52.1%.

It can be seen that k-means speaker clustering is greatly affected by the selection of initial clustering centers. The instability of clustering results directly leads to the reduction of the average clustering accuracy.

3.4.2. Speaker Clustering Based on Self-Organizing Neural Network. The number of neurons in the input layer is 96, the number of neurons in the competition layer is 3, the number of iterations is 500, and the learning rate is $\eta(t) = 0.1$. The self-organizing neural network speaker clustering algorithm described in Section 3.4.2 is used to cluster the MFCC feature set X_i $(i = 1, 2, \ldots, 96)$ (Table 3).

It can be seen that the accuracy of the self-organizing neural network algorithm is lower than the k-means algorithm when the initial clustering centers are selected appropriately.

However, because of its stable clustering results, the average clustering accuracy is higher than the k-means algorithm. Therefore, we try to combine the two algorithms and use the self-organizing neural network to improve the k-means algorithm, so that the clustering results are stable and can ensure a higher accuracy.

3.4.3. Improved k-Means Speaker Clustering Based on Self-Organizing Neural Network. MFCC feature set X_i $(i = 1, 2, \ldots, 96)$ is clustered by the improved k-means speaker clustering algorithm based on the self-organizing neural network.

First, the number of categories is predicted: assuming that the number of speakers in audio is unknown, the self-organizing neural network is used to predict the number of speakers. Let the number of classes be $k \leq 9$, 9 neurons are set up in the competition layer of the network, and the layout is 3×3. The number of neurons in the input layer is 96, and a small number of iterations (100) is set.

After training, count the winning times P of each neuron in the competition layer. Calculating 4/3 of the average number of wins is 14.22. There are three neurons with more than 14.22 wins, and the number of wins is 22, 20 and 18, respectively. It shows that the three neurons are closer to the actual clustering center, while the neurons with less than 14.22 wins are far away from the actual clustering center, which can be ignored. Therefore, the number of predicted categories is $k = 3$. (In order to accurately predict the number of categories, the mode number can be obtained by multiple discriminations.)

After predicting that the number of speakers is three, the self-organizing neural network is retrained. At this time, it is changed to set 3 neurons in the competition layer, and other things remain unchanged. At the end of the network training, the weight value $\{W_1, W_2, W_3\}$ of each neuron is obtained as shown in Figure 20.

	1	2	3	4	5	6	7	8	9	10	11	12
1	−0.2318	−0.6216	−0.8246	0.8197	−1	0.0115	0.0276	−0.7478	−0.3292	−1	−0.0117	−0.3540
2	−0.4504	−0.4722	−0.2641	0.6031	−0.4363	0.0251	−0.0120	−0.3220	−0.2556	−0.3009	−0.0195	−0.5682
3	0.0611	0.4078	0.4520	0.9593	−0.1953	0.2274	0.2702	0.4878	−0.1718	0.5584	−0.1076	−0.1543
4	0.0127	0.6880	−0.1491	−0.3808	−0.0066	0.0899	0.2659	0.5310	−0.4424	0.2297	−0.4673	−0.0847
5	0.1407	0.7320	0.4009	0.6745	0.1440	0.1724	0.0330	0.4351	−0.2905	0.5446	0.2133	−0.2824
6	−0.7343	−0.4021	−0.4176	−0.1656	−0.5628	−0.1922	−0/0962	−0.5104	−0.5630	−0.3270	−0.7263	−0.5573
7	−0.3879	0.7239	−0.1195	−0.3951	−0.3567	0.3155	−0.0020	0.6301	−0.0082	0.4713	−0.0835	0.0816
8	−0.9601	−0.7837	−0.3593	0.0225	−0.2102	0.0413	−0.1888	−0.6212	−0.4318	−0.3511	−0.1205	−0.2959
9	−0.8367	−1	−0.0377	0.4291	−0.6289	−0.1104	−0.4750	−0.9491	−0.3937	0.0958	−0.5412	−0.2736
10	−0.0155	0.2727	0.2820	0.6068	−0.1523	0.4739	0.3752	0.2055	−0.5360	0.1430	0.0281	−0.2739
11	−0.5196	−0.6370	−0.3923	0.1216	−0.5501	−0.0138	−0.2554	−0.5492	−0.3250	−0.4713	0.0153	−0.0793
12	0.2645	0.5404	−0.5999	0.3577	0.3054	−0.3061	0.2822	0.8176	−0.2828	0.6644	−0.4664	−1
13	0.1543	0.6268	−0.0959	−0.0056	−0.0863	−0.1137	0.3584	0.8135	−0.7636	0.2058	−3.0007	−0.5701
14	−0.3113	0.7890	0.1734	−0.0671	−0.6751	−0.1598	0.6133	0.5194	−0.4182	1	−0.3457	−0.5522
15	0.1356	0.7456	−0.1363	−0.1853	0.4774	−0.2183	0.0697	0.8824	−0.9199	0.4200	−0.8365	−0.3673
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
88	−0.5569	0.2714	−0.8230	0.2283	−0.0786	−0.7724	−0.2758	−0.2648	−0.6207	−0.3815	−0.2345	0.1627
89	−0.2697	−0.2357	−0.7891	−0.0462	0.3934	−0.3903	−0.1434	−0.5120	−0.0433	−0.8132	0.0258	−0.0983
90	0.0043	−0.0438	−0.8671	0.1035	−0.2169	−0.9042	−0.2299	−0.4181	−0.5621	−0.4401	−0.3949	0.1063
91	0.9374	−0.3438	−0.9019	0.2494	−0.9633	−1	0.4109	0.0358	−0.2700	−0.9540	−0.6541	−0.5618
92	0.4810	0.5058	0.5543	0.4640	−0.0819	0.2065	0.2168	9.4952e	0.1676	0.1380	0.0483	0.0595
93	0.2873	−0.2742	−0.2113	0.5668	0.0048	−0.2358	−0.2286	−0.0165	−0.2277	−0.0424	0.0560	0.0929
94	−0.2994	0.0670	−0.6855	0.0121	−0.2985	−0.6471	0.0806	−0.2228	−0.5288	−0.1501	−0.1191	0.6426
95	−0.6023	0.0626	−0.4080	−0.2971	−0.0838	−0.7251	−0.1777	−0.3717	−0.0202	−0.6585	−0.3076	0.5214
96	0.0239	−0.4341	−0.7470	0.2458	0.0398	−0.7044	−0.0350	−0.4216	0.0658	−0.4074	−0.4117	−0.0400

Figure 19: MFCC feature set.

Table 1: Speech category table.

Laber	1	2	3	4	5	6	7	8	9	10
Categories	a	a	b	c	b	a	c	a	a	b
Laber	11	12	13	14	15	16	17	18	19	20
Categories	a	c	c	c	c	c	b	b	a	c
Laber	21	22	23	24	25	26	27	28	29	30
Categories	c	b	a	a	a	a	b	a	a	a
Laber	31	32	33	34	35	36	37	38	39	40
Categories	a	a	b	c	c	c	c	c	c	b
Laber	41	42	43	44	45	46	47	48	49	50
Categories	a	a	a	a	b	b	a	b	c	a
Laber	51	52	53	54	55	56	57	58	59	60
Categories	a	b	c	a	a	c	c	a	a	a
Laber	61	62	63	64	65	66	67	68	69	70
Categories	c	c	c	b	b	b	a	a	a	a
Laber	71	72	73	74	75	76	77	78	79	80
Categories	c	c	c	c	c	a	a	a	c	c
Laber	81	82	83	84	85	86	87	88	89	90
Categories	a	c	c	b	c	c	a	a	a	a
Laber	91	92	93	94	95	96				
Categories	a	b	a	a	a	a				

Table 2: *K*-means speaker clustering results with improper initial values.

Laber	1	2	3	4	5	6	7	8	9	10
Categories	a	a	b	b×	b	a	b×	a	a	b
Laber	11	12	13	14	15	16	17	18	19	20
Categories	a	b×	b×	c	c	b×	b	b	a	b×
Laber	21	22	23	24	25	26	27	28	29	30
Categories	b×	b	a	a	a	a	b	a	a	a
Laber	31	32	33	34	35	36	37	38	39	40
Categories	a	a	b	b×	b×	c	c	b×	b×	b
Laber	41	42	43	44	45	46	47	48	49	50
Categories	a	a	a	c×	b	b	c×	b	b×	c×
Laber	51	52	53	54	55	56	57	58	59	60
Categories	c×	b	b×	c×	a	b×	b×	a	c×	c×
Laber	61	62	63	64	65	66	67	68	69	70
Categories	b×	b×	b×	b	b	b	c×	c×	c×	c×
Laber	71	72	73	74	75	76	77	78	79	80
Categories	b×	b×	b×	b×	b×	a	ca	a×	b×	b×
Laber	81	82	83	84	85	86	87	88	89	90
Categories	a	b×	b×	b	b×	b×	c×	c×	a	c×
Laber	91	92	93	94	95	96				
Categories	a	b	a	c×	c×	a				

Table 3: SOM speaker clustering results.

Laber	1	2	3	4	5	6	7	8	9	10
Categories	a	a	b	c	b	a	b×	a	a	b
Laber	11	12	13	14	15	16	17	18	19	20
Categories	a	c	c	b×	c	b×	b	b	a	c
Laber	21	22	23	24	25	26	27	28	29	30
Categories	c	b	a	a	a	a	b	a	a	a
Laber	31	32	33	34	35	36	37	38	39	40
Categories	a	a	b	c	a×	c	a×	c	c	b
Laber	41	42	43	44	45	46	47	48	49	50
Categories	a	a	a	b×	b	b	b×	b	c	a
Laber	51	52	53	54	55	56	57	58	59	60
Categories	a	b	c	a	a	c	c	a	a	a
Laber	61	62	63	64	65	66	67	68	69	70
Categories	c	a×	a×	b	b	b	a	a	a	a
Laber	71	72	73	74	75	76	77	78	79	80
Categories	c	c	a×	c	c	a	a	b×	c	c
Laber	81	82	83	84	85	86	87	88	89	90
Categories	a	c	c	b	c	c	a	a	a	a
Laber	91	92	93	94	95	96				
Categories	a	b	a	a	a	a				

	1	2	3	4	5	6	7	8	9	10	11	12
1	−0.4599	−0.2687	−0.4762	0.0375	−0.3098	−0.4637	−0.3003	−0.4836	−0.4448	−0.4555	−0.1798	0.0254
2	0.1069	0.5795	0.2901	0.4029	−0.0095	0.2195	0.1038	0.2853	−0.0727	0.2792	0.1259	−0.2594
3	0.3370	0.6511	−0.3395	−0.1476	0.1696	−0.2267	0.4405	0.5404	−0.4869	0.4999	−0.5759	−0.1712

Figure 20: Weights of neurons after training.

Table 4: SOM + k-means speaker clustering results.

Laber	1	2	3	4	5	6	7	8	9	10
Categories	a	a	b	c	b	a	c	a	a	b
Laber	11	12	13	14	15	16	17	18	19	20
Categories	a	c	c	c	c	c	b	b	a	c
Laber	21	22	23	24	25	26	27	28	29	30
Categories	c	b	a	a	a	a	b	a	a	a
Laber	31	32	33	34	35	36	37	38	39	40
Categories	a	a	b	b×	c	c	c	b×	c	b
Laber	41	42	43	44	45	46	47	48	49	50
Categories	a	a	a	a	b	b	a	b	c	a
Laber	51	52	53	54	55	56	57	58	59	60
Categories	c×	b	c	a	a	b×	c	a	a	a
Laber	61	62	63	64	65	66	67	68	69	70
Categories	c	c	c	b	b	b	a	a	a	c×
Laber	71	72	73	74	75	76	77	78	79	80
Categories	c	c	c	c	c	a	a	a	c	c
Laber	81	82	83	84	85	86	87	88	89	90
Categories	a	c	c	b	c	c	a	a	a	a
Laber	91	92	93	94	95	96				
Categories	a	b	a	a	a	a				

In Figure 20, the three rows of the matrix correspond to the values of W_1, W_2, W_3. The weight value $\{W_1, W_2, W_3\}$ is saved as the initial clustering center of the k-means algorithm.

Finally, the initial clustering center of the k-means algorithm is set as follows: $\mu_j = W_j$ $(j = 1, 2, 3)$ The implementation of the k-means speaker clustering algorithm and the end of the experiment are shown (Table 4).

To sum up, for Recording 1. wav, the improved k-means speaker clustering algorithm based on self-organizing neural network has achieved good clustering results. It effectively

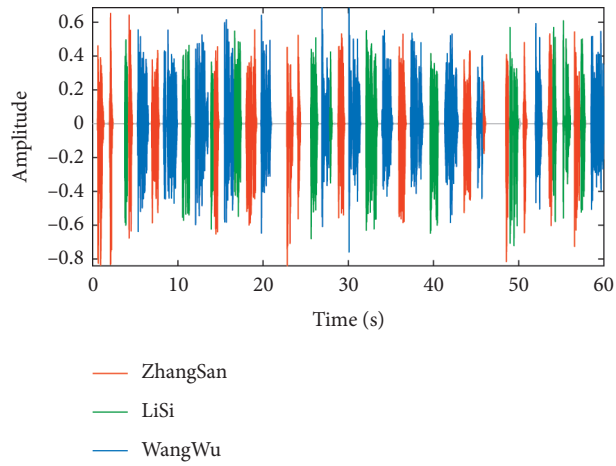

Figure 21: Diagram of SOM + k-means speaker clustering results. For clarity, a partial enlargement is shown in Figure 22.

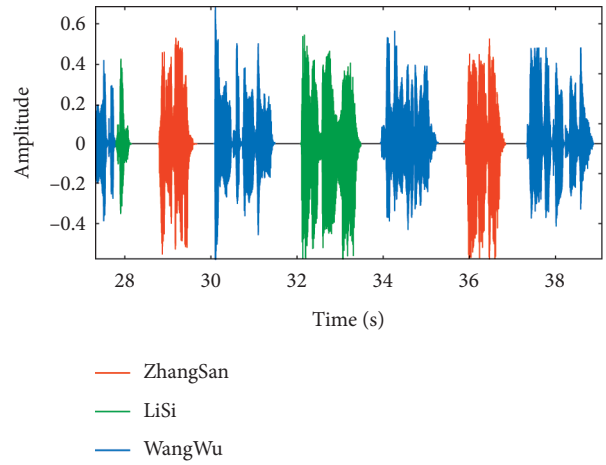

Figure 22: Diagram of SOM + k-means speaker clustering results (local magnification).

makes up for the shortcomings of the self-organizing neural network algorithm and k-means algorithm.

The clustering effect is shown in Figure 21, which distinguishes different speakers by different colors, and the image is intuitive. Among them, Zhang San's voice is red, Li Si's voice is blue, Wang Wu's voice is green, and the mute segment is gray.

3.4.4. Comparative Analysis. Nine audio files are selected to verify and analyze the above experimental results, added audio file Recording 1. wav, a total of 10 recordings. The contents of each audio file are as follows:

Recording 1: contains three voices, two men and one woman

Recording 2: contains two voices, two men

Recording 3: contains two voices, two women

Recording 4: contains two voices, one male and one female

Recording 5: contains three voices, one male and two female

Recording 6: contains three voices, three men

Recording 7: contains three voices, three women

Recording 8: contains four voices, two men and two women

Recording 9: contains four voices, two men and two women

Recording 10: contains four voices, three men and one woman

Based on the improved double-threshold endpoint detection method in this paper, three algorithms are used to perform speaker clustering experiments. The accuracy of each algorithm is shown in Table 5.

It can be seen from the table that the clustering accuracy of self-organizing neural network algorithm is low, but the average accuracy of k-means algorithm is often lower than that of self-organizing neural network algorithm because of its instability. With the increase of the number of speakers in the audio samples, or the decrease of gender differences, the clustering accuracy has a downward trend. However, in the same audio samples, the clustering accuracy

Table 5: Comparison of speaker clustering results based on three algorithms.

	K-means speaker clustering (%)	SOM speaker clustering (%)	Improved *k*-means speaker clustering based on SOM (%)
Sound recording 1	84.5	88.5	94.8
Sound recording 2	84.3	89.2	95.1
Sound recording 3	85.2	83.7	94.9
Sound recording 4	86.0	90.2	96.1
Sound recording 5	82.2	85.5	93.6
Sound recording 6	81.0	82.2	90.2
Sound recording 7	82.0	81.5	89.8
Sound recording 8	74.8	77.2	85.5
Sound recording 9	73.8	76.8	86.0
Sound recording 10	73.3	78.0	84.8

of the improved *k*-means algorithm based on the self-organizing neural network is always higher than the other two algorithms.

To sum up, compared with the *k*-means speaker clustering algorithm, the improved algorithm can not only predict the number of categories but also select the initial clustering center reasonably, so that the clustering results are stable. Compared with the self-organizing neural network speaker clustering algorithm, the improved algorithm reduces the number of iterations of the network, makes convergence faster, and greatly improves the clustering accuracy. Therefore, the improved *k*-means speaker clustering algorithm based on the self-organizing neural network is better than the self-organizing neural network algorithm and *k*-means algorithm.

4. Conclusion

The improved speech endpoint detection algorithm proposed in this paper can effectively eliminate the isolated noise points and enhance the antinoise performance of the algorithm. The threshold value is selected by the local maximum of the histogram of the statistical feature sequence, which improves the accuracy of speech detection. It enhances the ability of antinoise and meets the requirements of speaker segmentation better. Through the comparative analysis of the clustering accuracy of 10 recordings, it can be seen that with the increase of the number of speakers in the audio samples, the clustering accuracy of *k*-means and self-organizing neural network algorithms both decrease to 80%. However, the clustering accuracy of the improved *k*-means algorithm based on the self-organizing neural network is still maintained at 85%–89%. The improved *k*-means speaker clustering algorithm based on the self-organizing neural network improves the clustering accuracy, which not only makes up for the defects of the self-organizing neural network algorithm that the convergence is slow and cannot provide accurate clustering information, but also makes up for the defects of the *k*-means algorithm that the number of clusters needs to be given in advance and is greatly affected by the selection of initial clustering centers.

Data Availability

All of the data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the Natural Science Foundation of Liaoning Province (2019-ZD-0168 and 2020-KF-12-11), Major Training Program of Criminal Investigation Police University of China (3242019010), and Key Research and Development Projects of the Ministry of Science and Technology (2017YFC0821005).

References

[1] J. Yang, Z. P. Li, and P. Su, "Review of speech segmentation and endpoint detection," *Journal of Computer Applications*, vol. 40, no. 1, pp. 1–7, 2020.

[2] Q. Fan, "*Implementation and Performance Research of Speaker Logging System*", Ph.D. Beijing Normal University, Beijing, China, 2011.

[3] D. Z. Yang, J. M. Xu, J. Liu et al., "Reliable mute model and speech activity detection in speaker logs," *Journal of Zhejiang University (Engineering)*, vol. 50, no. 1, pp. 151–157, 2016.

[4] I. K. Sethi, "Video classification using speaker identification," *SPIE*, vol. 3022, pp. 218–225, 1997.

[5] F. Zheng, L. T. Li, and H. Zhang, "Voiceprint recognition technology and its application status," *Information Security Research*, vol. 2, no. 1, pp. 44–57, 2016.

[6] X. K. Li, Y. L. Zheng, N. Yuan et al., "Research on voiceprint recognition method based on deep learning," *Journal of Engineering of Heilongjiang University*, vol. 9, no. 1, pp. 64–70, 2018.

[7] A. Hannun, C. Case, J. Casper et al., "Deep speech: scaling up end-to-end speech recognition," *Computer Science*, vol. 17, pp. 1–12, 2014.

[8] H. Z. Chen and Z. J. Zhang, "A speech endpoint detection method based on energy and frequency band variance," *Science Technology and Engineering*, vol. 19, no. 26, pp. 249–254, 2019.

[9] N. Seman, Z. Abu Bakar, and N. Abu Bakar, "An evaluation of endpoint detection measures for Malay speech recognition of an isolated words," in *Proceedings of the 2010 International*

Symposium on Information Technology, vol. 10, pp. 1628–1635, Kuala Lumpur, Malaysia, June 2010.

[10] S. Morita, M. Unoki, X. Lu, and M. Akagi, "Robust voice activity detection based on concept of modulation transfer function in noisy reverberant environments," *Journal of Signal Processing Systems*, vol. 82, no. 2, pp. 163–173, 2016.

[11] Y. Zheng and S. Gao, "Speech endpoint detection based on fractal dimension with adaptive threshold," *Journal of Northeastern University (Natural Science)*, vol. 41, no. 1, pp. 7–11, 2020.

[12] W. U. Di, H. Zhao, C. Huang et al., "Speech endpoint detection in low-SNRs environment based on perception spectrogram structure boundary parameter," *Journal of Signal Processing Systems*, vol. 39, no. 4, pp. 392–399, 2014.

[13] M. Eshaghi and M. R. Karami Mollaei, "Voice activity detection based on using wavelet packet," *Digital Signal Processing*, vol. 20, no. 4, pp. 1102–1115, 2010.

[14] Y. Y. Lu, N. Zhou, K. Xiao et al., "Improved speech endpoint detection algorithm in strong noise environment," *Journal of Computer Applications*, vol. 34, no. 5, pp. 1386–1390, 2014.

[15] J. T. Liu and N. Jiang, "Research on speech segmentation and clustering based on mixed features," *Electro-Optic Technology Application*, vol. 34, no. 5, pp. 37–41, 2019.

[16] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 303–330, 2006.

[17] G. R. Hu and X. D. Wei, "Endpoint detection of noisy speech based on cepstrum feature," *Journal of Electronics*, vol. 28, no. 10, pp. 95–97, 2000.

[18] L. Li and J. Zhu, "Research on speech endpoint detection based on wavelet analysis and neural network," *Journal of Electronic Measurement and Instrument*, vol. 27, no. 6, pp. 528–534, 2013.

[19] P. Delacourt and C. J. Wellekens, "DISTBIC: a speaker-based segmentation for audio data indexing," *Speech Communication*, vol. 32, no. 1-2, pp. 111–126, 2000.

[20] Z. P. Zhang, L. N. Zhang, and S. He, "Research on continuous adaptive algorithm based on GMM-UBM speaker model," *Communication Power Supply Technology*, vol. 33, no. 2, pp. 81–83, 2016.

[21] C. B. Huo, C. J. Zhang, and H. M. Zhao, "Research on speaker verification system based on GMM-UBM," *Journal of Liaoning University of Technology: Natural Science Edition*, vol. 3, pp. 149–151, 2012.

[22] B. Fergani, M. Davy, and A. Houacine, "Speaker diarization using one-class support vector machines," *Speech Communication*, vol. 50, no. 5, pp. 355–365, 2008.

[23] W. X. Zhu, "*Research on Speaker Segmentation and Clustering in Multi-Person Conversation Scene*", Ph.D. University of Science and Technology of China, Hefei, China, 2017.

[24] H. Qiu, "*Research on Speaker Clustering Based on GMM and Hierarchical Clustering*", Peking University, Beijing, China, 2004.

[25] J. L. Ma, X. X. Jing, and H. Y. Yang, "Application of principal component analysis and *K*-means clustering in speaker recognition," *Computer Application*, vol. 35, no. s1, pp. 127–129, 2015.