Research on Tibetan Speech Endpoint Detection Method Based on Extreme Learning Machine

Ze-guo Liu(✉)

Key Lab of China's National Linguistic Information Technology, Northwest Minzu University, Lanzhou 730000, China

Abstract. The traditional method of Tibetan speech endpoint detection will reduce the detection accuracy in low SNR environment, so a new method based on limit learning machine is proposed. Firstly, speech signal is preprocessed. By analyzing the human discourse generation model, the speech signal is filtered and processed by frame, the high frequency interference signal in the signal flow is filtered, and the signal flow is decomposed into multiple frames by using the short-term stationary characteristics of the speech signal, so that the characteristics of the speech signals in each frame are kept constant; the sentence analysis is optimized and the word association in the sentence is analyzed by using line graph syntax The system forms an independent semantic block, introduces the pattern template based on pseudo matrix, generates pseudo points through the original feature vector structure and inserts it into the original speech features for protection. Finally, the classification algorithm of limit learning machine is introduced, the optimal configuration of i-h-o is selected, and the end detection method of Tibetan speech based on limit learning machine is completed. The simulation results show that the accuracy of the speech endpoint detection results is significantly higher than that of the traditional method under different SNR and ambient noise.

Keywords: Limit learning machine · Tibetan speech · Speech endpoint detection

1 Introduction

Tibetan has always played an important role in the past and present. It is an indispensable tool for Tibetan compatriots to communicate with the outside world, and also the carrier of Tibetan culture, recording and inheriting the splendid historical culture. With the development of society, ethnic culture is more and more inclusive, and minority characters are gradually loved by people. Tibetan language, as an integral part of minority languages, has also played an important role in the international world, not only by the recognition of international standard words, but also by the recognition of the global highway [1, 2]. Therefore, whether from the actual development of the country or from the perspective of cultural inheritance, Tibetan studies have very important significance. The structure of Tibetan mainly depends on syllables to realize the expression of words and sentences. The phonetic segment is closely related to syllable. When the

S. Liu and X. Ma (Eds.): ADHIP 2021, LNICST 416, pp. 480–490, 2022.
https://doi.org/10.1007/978-3-030-94551-0_38

phonetic segment is found, the end point of syllable is easy to find. The study on the characteristics of Tibetan speech is the protection of Tibetan culture and helps to better use Tibetan characters. In the process of speech detection, the fluctuation of vocal cord appears gradually, which makes the whole syllable promote the implementation of the detection. When the syllable appears the main vowel, the audio is the maximum. At this time, the speech components can be analyzed according to the language characteristics of Tibetan language, and the speech can be detected according to the time frequency or the frequency of the voice vibration. Analogy is the process of hearing detection in the natural noise environment. Therefore, it can achieve the purpose of separating Tibetan speech signals. As far as speech signal is concerned, the clear part and the voiced part can be distinguished from the spectrum characteristics of speech. The spectrum of the former voice is rising at more than 4 kHz, while the latter is declining and above 4 kHz.

With the continuous progress of society, great changes have taken place in the way of daily communication. Voice communication is an important form of people's emotional interaction. At the same time, speech signal processing technology has developed rapidly. The purpose of speech signal processing is to determine the starting and ending point of speech signal in the presence of noise. When the effect of endpoint detection is good, the first thing we can see is that the processing time decreases significantly, and then there will be noise interference, which needs to be eliminated by certain technology, so as to increase the performance of the detection system. There are many endpoint detection methods, among which the short-term energy method was invented in the earliest period. With the passage of time and the maturity of the theory, the form of zero crossing rate method appeared again. The premise of these methods to achieve good results is to carry out experiments in the environment of high signal-to-noise ratio. When the environment becomes the environment of natural noise and low signal-to-noise ratio, its effect will be significantly weakened. Reference [3] combines spectral subtraction with energy zero ratio method. With the help of spectral subtraction, the signal-to-noise ratio of speech signal is improved, and the endpoint detection of processed speech is carried out by energy zero ratio method. Finally, the experimental analysis of Tibetan speech is needed. However, the accuracy of speech endpoint detection result of this method is low.

In view of the problems of the above methods, this paper proposes a Tibetan speech endpoint detection method based on extreme learning machine, and verifies the effectiveness of this method through simulation experiments, which solves the problems of traditional methods.

2 Research on Tibetan Speech Endpoint Detection Method Based on Extreme Learning Machine

2.1 Speech Signal Preprocessing

Speech signal is a kind of nonlinear time-varying and discrete complex signal. From the whole, the speech signal is a nonstationary signal. But the local analysis shows that the signal has the short-term stability characteristic when the signal is consistent in about 30 ms. To realize the analysis of the signal, first of all, we should understand the sound

generating principle of the voice signal. The voice signal is an acoustic signal sent out by the cooperation of multiple organs of the human body. It has high frequency part and low frequency part, and also has the clear sound and the voiced component [4]. According to these characteristics, we will preprocess the language signal. Human voice is produced by the coordination of multiple functional organs. The speech signal is produced by these processes. When the direct air flow from the lungs flows through the complex and changeable channels composed of trachea, throat, upper jaw and lip, different sounds are produced due to the changing channels. In short, what kind of speech to send out is determined by the combination of the tongue, upper jaw and lip. Third, the radiation system, through the above two systems, has been produced, and it needs to be radiated

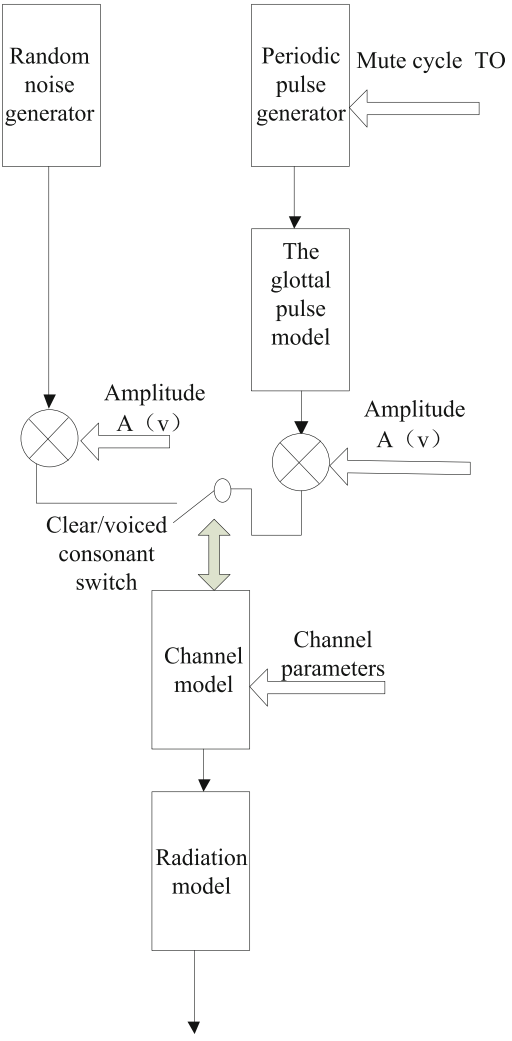

Fig. 1. Speech signal generation model

outside the lip and nasal cavity. So the human discourse we hear is just like this. The specific model diagram is as follows (Fig. 1):

The purpose of speech signal filtering is to filter out the high-frequency interference signal in the signal stream, which is usually bounded by half of the sampling frequency f_S, and the band larger than that frequency will be isolated [5, 6]. The reason why the filter is set as a band-pass filter is that the 50 Hz power frequency interference must be filtered at the same time. f_H, f_L will be set as the upper and lower limit of the filter frequency. Generally speaking, the upper and lower limit frequencies are $f_H = 3.44$ Hz, $f_L = 60$–100 Hz, and f_S is usually 8 kHz. For the design of speech recognition, the upper and lower limits of intermediate frequency rate are $f_H = 4.5$ Hz or 8 kHz, $f_L = 60$ Hz, and the selection of f_S is usually 10 kHz. Then, according to the sampling theorem, f_S is selected, and the value of sampling frequency f_S must meet the standard of more than 2 times of the frequency of the sampled signal. A/D converter generally selects 8-bit or 12 bit, and then performs A/D conversion on speech signal.

After the completion of the filtering process, it also needs to carry out the frame processing. From the overall time-domain waveform of the speech signal, the signal characteristics are in a non-stationary state, but from the micro point of view, in a very short period of time, its characteristics are in a stationary state, almost unchanged, which we call the short-term stationary characteristics of the speech signal. This feature brings a breakthrough for the complex and changeable speech signal processing, which can decompose the signal flow into several small segments, that is, a frame, in which the characteristics of the speech signal remain constant [7, 8]. Therefore, the frame processing of speech signal is the beginning of speech analysis. The signal can be decomposed into multiple parts of a frame of 10 ms–30 ms, and then each frame of speech signal can be analyzed as the research object. However, if the whole speech information is simply cut into multiple speech segments, the feature parameters between the adjacent frames will change too much, which is not conducive to signal analysis and processing.

In order to make the speech feature of adjacent frames have the characteristics of smooth transition, the signal of the next frame is often moved forward for a distance, called frame shifting, which makes the adjacent frames cross overlap, so that the feature parameters of adjacent frames have relative consistency [9, 10]. In the design of speech recognition, the length N of each frame is generally within 10 ms–30 ms, while the size of frame shift M is generally more than 1/3 of frame length N. It is important to select the frame length. If it is too long, it violates the short-term stability of speech signal, which makes the feature parameters of a frame different and can not find the correct feature value. If the selection is too short and the number of frames is too many, the calculation amount will be greatly increased and the operation time will be prolonged. At this point, the speech signal preprocessing is completed.

2.2 Analysis of Optimized Sentences

In the original endpoint detection method, the detection of simple phrases is relatively easy, but multiple backtracking will affect the efficiency of the algorithm, so this paper chooses the line graph syntax. The algorithm involves three data structures: line graph, process table and active edge set. These three data structures include more information.

In the analysis process of the algorithm, when the process table is blank, in order to obtain the part of speech tags of the next language input, the start and end positions will be stored in the process table together, and an element will be extracted from the process table and recorded as $X(i, i')$, where i represents the start position, i' represents the end position, and according to certain rules, the rule set will not be changed The matching points are marked and added to the active edge set together with the extracted element $X(i, i')$, and then the extension subroutine is called. The analysis results are as follows (Fig. 2):

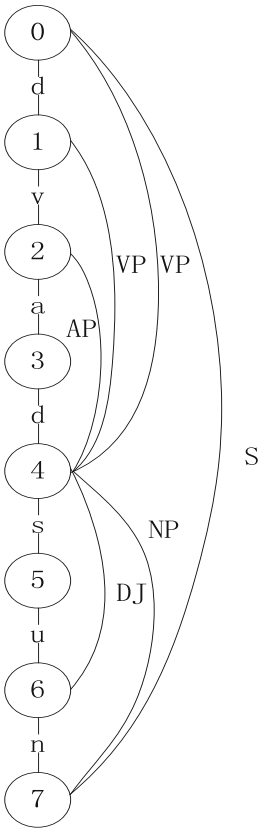

Fig. 2. Analysis result curve

The symbols in the figure above are marked with reference to the classification of symbols and the definitions in the marking Grammar Library. After using syntactic analysis and applying it to the original input string of Tibetan speech, we can get the algorithm result graph as shown in the figure above. From the result graph, we can see the relationship and connection between words in sentences, and further form independent semantic blocks on the basis of these relationships. After morphology and analysis, new rewriting rules can be obtained. The original input sentences of Tibetan speech can be processed to get a limited number of semantic blocks. After processing, these

semantic blocks show the characteristics of speech endpoint, and they are identified and classified. In order to improve the recognition accuracy of this method, this paper introduces the voiceprint template based on the chaff matrix, generates the pseudo points (CPS) through the original feature vector elements, and inserts the CPS into the original speech features for protection. Compared with the existing system based on Fuzzy vault (FV), this method has stronger security. The feature extracted in the registration process is X_0, different authentication feature $X_I, X_J \in R^{m\times p}$, where m is the feature length of audio and p is the order of Mel cepstrum coefficient. The pseudo matrix of chaff matrix is expressed as C, $C \in R^{m\times r}$ and r, which indicate the number of CPS added. The feature of positioning matrix after adding C can be expressed as follows:

$$X_{C0}, X_{CI}, X_{CJ} \in R^{m\times(p+r)} \tag{1}$$

In the above formula, X_{C0} is the voiceprint template designed in this paper, which is composed of codebook inserting matrix C, and all audio features after adding C matrix are collectively referred to as X_C. In order to simulate the perception of human ear to sound in real situation, a sound signal scale converted into pseudo matrix is selected to reflect the frequency range of spectrum coefficient. Firstly, the linear spectrum of the sound frequency is selected by Fourier transform, then the transverse axis of natural spectrum is scaled nonlinearly, and a new energy spectrum in the cepstrum coefficient of Mel frequency is obtained. The Mel frequency cepstrum is obtained by the inverse spectrum analysis.

The linear transformation of horizontal spectrum axis is to convert the speech spectrum into a spectrum range. Mel spectrum coefficient module is used to convert the audio spectrum module into an audio spectrum range. Then the transfer function of each triangle filter is as follows:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1) \le k \le f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)}, & f(m) \le k \le f(m+1) \\ 0, & f > f(m+1) \end{cases} \tag{2}$$

In the above formula, $1 \le m \le M$ and M represent the number of triangular filters, and $f(m)$ represents the center frequency of the m-th filter. Through the filter, the characteristics of speech can be obtained from the linear natural spectrum to the nonlinear Mel spectrum, so that a frame of speech sample can be represented by a 22 dimensional vector. It can be used as input to the back-end identification part of the system.

2.3 Introducing the Classification Algorithm of Limit Learning Machine

Extreme learning machine (ELM) has been effectively applied in different pattern recognition applications. Many applications show that extreme learning machine has better accuracy and lower time consumption than SVM in small data, sparse data to medium large data. Therefore, elm is used as Tibetan speech endpoint detection classifier in this paper. Elm is a variant of artificial neural network. It can solve the problem of too long training time by randomly selecting hidden nodes and analyzing the weight of output

layer. Therefore, it is of great significance to quickly solve the regression and classification model. Elm has two important parameters, the number of hidden nodes and the regularization coefficient. The default number of neurons in the output layer is equal to the number of classes, and the regularization coefficient can also be changed. The standard range is 0.1–10. The function of elm mainly includes two steps: feature mapping and elm learning. For training sample set $\{x_i, t_i\}, i = 1 \ldots, N$, there are a total of N samples. Each sample x_i is a d dimensional column vector, and t_i is the output label. In the feature mapping stage, the input is transformed into a hidden layer, and the output function of elm is as follows:

$$f_L(x) = \sum_{j=1}^{L} \beta_j G(a_j, b_j, x_i) = h(x_i)\beta \tag{3}$$

In the above formula, $\beta = [\beta_1 ... \beta_L]^T$ represents the output weight vector between the hidden layer (L nodes) and the output layer:

$$h(x_i) = [h_1(x) ... h_L(x)]^T \tag{4}$$

In the above formula, the function of $h(x_i)$ is that x_i maps the d dimensional space to the L dimensional feature space, $G(a, b, x)$ is a continuous piecewise nonlinear function, and (a_i, b_i) is the parameter of the ith hidden node:

$$G(a, b, x) = \frac{1}{1 + \exp(-(a \cdot x) + b)} \tag{5}$$

In the above formula, parameter a, b is generated randomly. In the learning stage, the main goal is to obtain the minimum training error and norm output weight β. the general approximate values are as follows:

$$\lim_{L \to \infty} \left\| \sum_{i=1}^{L} \beta_j h_j(x) - f_L(x) = 0 \right\| \tag{6}$$

This paper uses i-h-o configuration for elm, which corresponds to input layer, hidden layer and output layer respectively. The number of neurons in I is equal to the number of input features, the number of neurons in O is the number of output classes (in this paper, there are two, that is, speech frames and noise frames), and the number of neurons in H is between 50 and 600. Based on the experiment, the most appropriate value is 360. So far, the research of Tibetan speech endpoint detection method based on extreme learning machine is completed.

3 Simulation Experiment

3.1 Setting Simulation Experiment Parameters

In order to extract the speech signal characteristics required by limit learning machine, the software mainly uses MATLAB r2014a. All experiments are carried out in Windows 10

professional edition. Core (TM) i5-7300hq processor is used, and the physical memory is 16 GB. By comparing the results of traditional Tibetan speech detection, the validity and performance of the method based on limit learning machine are determined. In the parameter setting of simulation experiment, the performance of the endpoint detection system is tested by using true positive rate (TPR), false positive rate (FPR) and the time and precision parameters of feature extraction, modeling, training and testing of classifier.

Endpoint detection is a binary classification problem, because any frame either contains speech signal or does not contain it. Assuming a given speech signal, the parameters of the classifier after classifying the frame are defined as follows:

$$\begin{aligned} TPR &= \frac{L_1}{N_1} \\ FPR &= \frac{L_2}{N_2} \\ ACC &= \frac{L_1+N_2-L_2}{N} \end{aligned} \tag{7}$$

In the above formula, L_1 represents the true positive parameter, L_2 is the false positive parameter, and ACC represents the detection accuracy. N represents the total number of voice frames, N_1 represents the total number of frames containing voice signals, and N_2 represents the total number of frames without speech signals, so there are:

$$N = N_1 + N_2 \tag{8}$$

In the data set of simulation experiment, TIMIT standard voice database and noise-92 standard noise database are used to train and test the VAD method proposed in this paper. In order to create a real simulation environment, the short sentences in the database are merged (the total time is 5–10 s) for simulation, in which the ratio of talking frame and non talking frame is kept between 1:3 and 3:1. The sentences in the dataset are divided into six different types of real-world noise: babble noise, street noise, room noise, restaurant noise, car noise and train noise. The signal-to-noise ratio is set to 0 dB, 5 dB, 10 dB and 15 dB. The feature vector of the frame is composed of low-frequency denoising energy, MFCC and formant frequency. In order to extract these features, the input speech signal is divided into frames with frame length of 25 ms and frame shift of 10 ms. In the experiment, *TPR*, *FPR* and *ACC* obtained by the trained extreme learning machine classifier are compared and analyzed with the existing detection methods, and the total time of the two detection methods in classification training and testing is calculated, and the results are compared and analyzed.

3.2 Result Analysis

Under the above experimental conditions, the true positive rate and false positive rate of this method under different SNR and noise types are obtained (Table 1).

The above table shows the test results of this method. It can be seen that *TPR* and *FPR* are not the same for different types of noise conditions and SNR levels, which indicates that the sensitivity of the improved features to different noises is not the same. Through calculation, the average *TPR* and *FPR* values of the two methods under different noise types are obtained as follows (Table 2):

Table 1. Statistical results of true positive rate and false positive rate of this method at 15 dB

Noise type	*TPR*	*FPR*
Babble	98.64	2.13
Street	96.25	0.58
Room	88.14	1.35
Restaurant	94.25	0.64
Car	98.64	0.00
Train	99.22	0.00

Table 2. Average *TPR* and *FPR* values of the two methods

Signal to noise ratio		Method	
		Method of this paper	Traditional method
15 dB	*TPR*	97.04	73.52
	FPR	0.84	10.36
10 dB	*TPR*	94.56	70.74
	FPR	1.32	13.59
5 dB	*TPR*	83.51	61.37
	FPR	1.58	20.35
0 dB	*TPR*	78.02	50.67
	FPR	3.11	28.31

The table above shows the comparison between the average *TPR* and *FPR* of the proposed method and the traditional method under six different noise types. The proposed method has obvious improvement compared with the traditional method.

The comparison of accuracy results is shown in the figure below (Fig. 3):

The figure above shows the average accuracy of the proposed method compared with the traditional method for endpoint detection. Under low SNR, the proposed method can greatly improve the accuracy of endpoint detection.

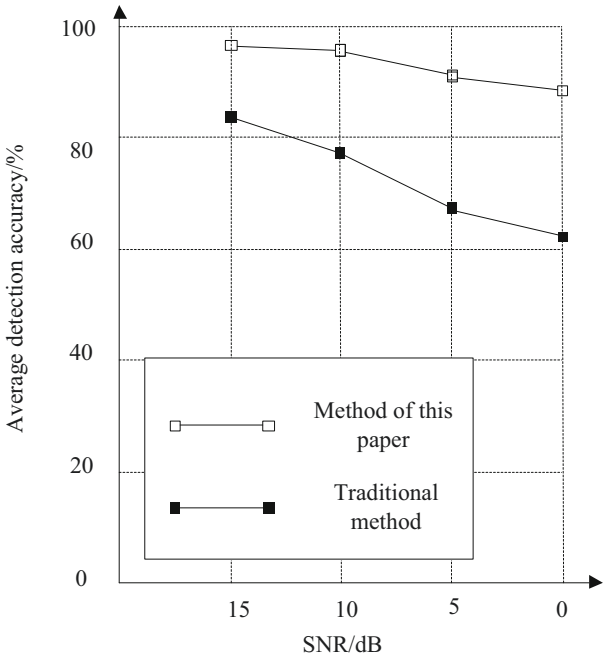

Fig. 3. Comparison of the accuracy of the two methods

4 Conclusion and Outlook

An excellent speech endpoint detection method can maintain high detection accuracy in the complex low SNR noise environment, and escort other steps in the speech recognition system. In this paper, the speech signal is taken as the research object, and the speech endpoint detection in low SNR environment is realized by integrating extreme learning machine and other methods. The simulation results show that the method designed in this paper is effective. Due to the limited research time and ability, the selection of noise signal in this paper is single noise, because of the lack of mixed noise library, such as white noise and pink noise mixed noise, car noise and human noise mixed noise. Therefore, in order to verify the endpoint detection effect of the proposed method in mixed noise, we need to enrich the noise database and supplement the corpus of various mixed noise.

Fund Projects. Funded by Graduate Research Innovation Project of Northwest Minzu University (Granted No. YXM2019010).

References

1. Zhang, T., Liu, Y., Ren, X.: Voice activity detection based on long-term power spectrum variability. J. Front. Comput. Sci. Technol. **13**(09), 1534–1542 (2019)
2. Xiangfeng, W., Yi, Y., Quan, Z., et al.: A dataset of Mongolian, Tibetan and Uyghur speech fragments based on voice activity detection. Sci. Data China (Chin. Engl. Online) **4**(04), 112–122 (2019)
3. Xie, X.: Research on endpoint detection method of Tibetan speech. Inf. Comput. (Theor. Ed.) **32**(450(08)), 106–108 (2020)
4. Zhou, B., Wei, S., Tang, Y., et al.: ELM network intrusion detection algorithm based on rough set attribute reduction. Transducer Microsyst. Technol. **38**(01), 122–125 (2019)
5. Yang, X., Ma, Z., Shen, H., et al.: Fault diagnosis of airflow jamming fault in double circulating fluidized bed based on multi-scale feature energy and KELM. CIESC J. **70**(07), 2616–2625 (2019)
6. Wei, J., Zhou, B., Tang, H., et al.: Transformer fault diagnosis with the combination of RapidMiner-modified particle swarm optimization-Extreme Learning Machine algorithm. Proc. CSU-EPSA **31**(03), 133–138 (2019)
7. Han, T., Zhang, H., Zheng, Z., et al.: Auditory perception speech signal endpoint feature detection based on temporal structure. J. Jilin Univ. (Eng. Technol. Ed.) **49**(01), 313–318 (2019)
8. Liu, S., He, T., Dai, J.: A survey of CRF algorithm based knowledge extraction of elementary mathematics in Chinese. Mob. Netw. Appl. https://doi.org/10.1007/s11036-020-01725-x
9. Liu, S., Pan, Z., Cheng, X.: A novel fast fractal image compression method based on distance clustering in high dimensional sphere surface. Fractals **25**(4), 1740004 (2017)
10. Liu, S., Fu, W., He, L., Zhou, J., Ma, M.: Distribution of primary additional errors in fractal encoding method. Multimedia Tools Appl. **76**(4), 5787–5802 (2014). https://doi.org/10.1007/s11042-014-2408-1