Research Status and Future Development of Endpoint Detection Algorithms Based on Computer Science Language Signals

Juan Wang¹,*
¹LUliang University, Computer Science and Technology, LUliang, Shanxi, China
*Corresponding author e-mail: wangjuan@llhc.edu.cn

Abstract. Today is an information technology era, language signal recognition technology, language signal coding technology and a variety of new language technology will be widely used in all areas of our life, such as: security field, human-computer interaction, communication field. We can accurately analyze pure speech signals and silent segments in a speech by using language signal endpoint detection technology, which will bring decisive effects to the efficiency of ASR and ASC. Three steps can be used to represent the language endpoint detection model, namely, the language signal preprocessing step, the extraction of the characteristic vector of the whole language flow, and the establishment of the language endpoint discrimination model. Finally, the speech endpoint discrimination model is established. The traditional speech endpoint detection algorithms include the two threshold method based on time domain, the universal entropy method based on frequency domain and the invert characteristic method. Aiming at the low SNR and complex noise environment, in order to obtain satisfactory endpoint detection effect, this paper proposes an endpoint detection model based on optimized Extreme Learning machine (ELM), and makes up for the deficiencies of the algorithm itself by optimizing network connection parameters[1-2].

Keywords: Speech Recognition, Speech Endpoint Detection, Preprocessing, Language Flow

1. Introduction
Speech is the main medium for many creatures to transmit information in nature, and it carries the most direct, explicit and clear information. With the progress of science and technology, signal processing technology provides powerful tools for the processing of speech information 8, and also pushes the research of speech signal analysis and processing to a higher position. It has become the research direction of many scholars[3-5].

1.1. Background and significance of speech endpoint detection
Language has been used to convey and express all kinds of information and instructions since the beginning of human society. Language is the driving force for social operation, ethnic integration and
the development of human civilization for thousands of years. Since modern times, the rapid development of information technology has opened up more and more channels of information transmission for human society[6-10]. However, no matter how human society develops and how science and technology progress, the way of using speech form to convey information is not weakened, but more widely used and plays a more obvious role in People's Daily life and commodity production. In recent years, the development of computer network technology and information technology is in full swing, and the growth momentum has been maintained in the forefront of each industry. Speech is not only the sound but also the meaning. It is the unique acoustic information carrier. In today's information age, research on how to integrate voice into modern industrial enterprises, home security, human-computer interaction and other fields will bring great convenience and innovation to the development of the entire industrial system and people's life. Such as today's industrial robots in production and processing, disaster relief, agricultural production and other fields has great application value and potential, by adopting what kind of new voice analysis technology, let the human discourse directly accepted by the robot, and according to the word meaning, so as to realize the great convenience of production and the safety of rescue and relief and efficient [10]. The computer has been popularized to every corner of human society's production and life, but when making instructions to the computer, the human-computer interaction mode we use is kept in the most primitive mouse, keyboard and so on, in some. There are still many inconveniences in this case. If some new speech information processing technology can make the computer understand human voice commands, by setting a distance voice transmission can still be accurate and complete receiving orders issued by human voice, can communicate accessible and human freedom, this will not only bring the revolutionary development of computer technology, will also promote the human society - a entered into a new stage!

1.2. The general process of speech recognition

Model is shown in Figure 1.

![Figure 1. Speech recognition.](image)

It can be seen from the description in Figure 1. that in THE ASR model, speech endpoint detection is the foundation and key link of the whole recognition model. The impact of endpoint detection on model performance can be shown in the following aspects. (1) In speech recognition or pronunciator recognition technology, if the influence brought by the channel is eliminated by using the inverted mean method, the mean value of the pure speech affected by the channel in the whole signal flow can be accurately calculated only when the speech frame is accurately marked and the non-speech frame
is discarded, so as to achieve the ideal recognition result. (2) In speech recognition technology, only by accurately distinguishing the pure speech and the mute segment can the interference of the mute segment noise and the integrity of the pure speech signal characteristic quantity be guaranteed, so as to ensure the accuracy of speech recognition. (3) When all kinds of noises are superimposed on the speech signal flow, the application environment becomes more complicated, and the difficulty of modeling will increase at this time. When the pure speech segment is accurately screened, the language information and pure noise can be modeled respectively. (4) Studies have shown that, in daily communication or normal communication, the pure speech segment only takes up a small part of the whole speech signal flow, which means that most of the speech signals are useless or even unfavorable in the analysis and processing of the speech information. When the mute segment is accurately distinguished, the mute segment information is removed from the whole information flow, which will greatly improve the speed of speech recognition. For complex noise environment, only accurate speech endpoint detection can provide reliable speech materials for subsequent speech processing.

1.3. Development status of voice endpoint detection technology
Speech endpoint detection technology is an important speech signal processing technology. As early as 1876 when Bell Laboratories invented the telephone, it began to study the speech signal processing technology. The invention realized the long-distance transmission of speech information for the first time through the transformation of acoustoelectric and electro-acoustic technologies. Since 1940, the emergence of vocoder and visible speech has successfully simulated the human speech system, and the model realizes the frequency of speech signals. Analysis of spectral features and parameters. However, apart from the laboratory environment and the appearance of various noises in practice, the analysis of speech information becomes complicated and difficult. At this time, how to find out the pure speech information in the speech signal and thus only process the pure speech information becomes a breakthrough in the research. So the voice endpoint detection technology arises at the historic moment.

After decades of development and improvement, a variety of endpoint detection algorithms have been emerging, some of them are outstanding in some environments, and some of them have excellent performance under the conditions. To sum up, speech endpoint detection can be roughly divided into the following categories: First, the earliest proposed methods based on time-domain characteristics are familiar and representative ones such as short-time energy method and double-threshold method. Their detection principle is relatively simple, the algorithm implementation complexity is also relatively low. The results are satisfactory in the laboratory environment, but the accuracy and stability of the method are greatly reduced in the noisy environment. - some scholars improved detection method based on energy, such as based on short-time linear energy method and based on the normalized energy method, the improved method to some extent, improved the miss rate and decreased rate of endpoint detection, and compared with the original algorithm is a kind of great progress, but in a complex environment, appear a lot less than the method based on time domain, even in low SNR environment error Secondly, the endpoint detection algorithm based on frequency domain is proposed. In order to change the defects of the time-domain endpoint detection method and improve the performance of the endpoint detection in the noisy environment, Japanese scholar Boyashi proposed in the 1980s a method to directly extract the frequency-domain information of speech signals based on the fast Fourier Transform (FFT), thus forming the frequency-domain endpoint detection algorithm. After that, endpoint detection in frequency domain has become the target of many people, and many frequency-domain based methods have been emerging. Such as spectrum variation characteristics, inherent energy characteristics, multi-band modulation energy characteristics, general correlation function, speech fundamental frequency information, and the minimummel-scalefrequency band parameters proposed by Lin et al., a Scholar from Taiwan, China. In the 1990s, the upentropy-based endpoint detection algorithm derived from the spectrum distribution of speech signal cosine transform was proposed. Subsequently, some improved
algorithms based on THE entropy are proposed one after another, and the more classical one is the weighted end point detection based on the entropy parameter weighting of short-term energy.

2. System model for detection
When detecting the starting and ending points of pure speech, it must be pretreated. Usually the type of signal we can access is analog signal, in order to meet the requirements of computer recognition, the analog signal must be converted to digital signal. The characteristics of the digitized speech signal can be represented by some characteristic parameters, but these characteristic parameters must represent the characteristics of the original signal as comprehensively as possible. In turn, these characteristic parameters as the input vector of the neural network into a neural network system, neural network system will value is calculated according to the training, simulation, adjustment of structural parameters, such as a series of changes, so that the output neurons can quickly and accurately output calculation results, the output value is the position of the speech signal endpoint, which is based on the neural network method of general process for speech endpoint detection. The specific flow chart is shown in Figure 2.

![Figure 2. Neural network endpoint detection simulation.](image)

According to Figure 2, we can clearly see that the process of the detection model is mainly divided into three steps, which are: convert the original speech signal into digital signal that can be recognized by the computer; The characteristic parameters of digital speech signal are analyzed. Establish the recognition and discrimination model, train and test the network model, and then output the detection results. It can be known that the specific process of these three steps is:

Step1: convert these non-wav files into the wav format required by the computer. The sampling frequency is generally 44100HZ, and the audio files processed by sampling and quantization can be directly run on data processing software such as Cool Edit Pro and matlab.

Step2: speech signal feature extraction, the so-called characteristic is characterized as comprehensively as possible speech signal data set, the characteristic of the information contained in the whole period of the voice stream, as well as pure speech signal and noise or noise signal flow characteristic of information is completely different, if we can put these information feature analysis, you can substitute for the analysis of the characteristics of the analysis of the original information. At present, there are many algorithms for extracting speech signal feature, such as wavelet analysis, empirical mode decomposition and so on.

Step3: Establish and test the neural network model, which is a discriminant model that can be used for speech endpoint detection. Usually, we take the extracted speech signal feature vector as the input of the neural network, and after the feature vector is input into the neural network, we repeatedly train the neural network with the training data set.

3. Conclusion
With the gradual application of speech recognition technology and speech coding technology in people's production and life, people have great expectations for the continuous improvement and wide spread of the technology. As the key link of the technology, speech endpoint detection has become the focus of many scholars' research. In this paper, the detection algorithm is introduced in detail, which is roughly divided into three parts. In the speech preprocessing part, the conventional speech filtering, framing and windowing methods are introduced. The theory of wavelet analysis and the
characteristics of multiresolution analysis are introduced in detail. The wavelet wavelet energy, its mean value and variance are extracted as the characteristic quantities of the frame speech signal. Finally, the speech endpoint discrimination model is established. This paper introduces the traditional two threshold method based on time domain, the entropy-based method based on frequency domain, the speech endpoint detection algorithm based on invert sign and the BP neural network algorithm, and points out the shortcomings of these algorithms.

References

[1] Zhang Junchang, Hu Haitao, Cui Li et al., speech endpoint detection integrating Burg spectrum estimation and signal change rate measure [J]. Journal of Xidian University (Natural Science), 2014, (3): 192-195.

[2] Chung Hoon Lee, Sung Joo Lee. Weighed - Finite State Transducer - -based Endpoint Detection Using Probabilistic DecisionLogic [J]. J ETRI JOURNAL, 2014, 36 (5): 714-720.

[3] He Ling, huang hua and so on speech endpoint detection Based on critical band and energy entropy [J]. Computer applications, 2013, 33(1).

[4] Baisen, Z. Ye and Z. Wulin, "Speech Endpoint detection with Low SNR Based or HHTSM [C]." Electronic Measurement & Instruments (ICEMI), 2013 IEEE 11th International Conference on Harbin, 2013: 116-119.

[5] Li Jie, Zhou Ping, Du Zhiran and other short-time TEO energy application in Speech Endpoint Detection with Noise [J]. Computer Engineering and Applications, 2013, 49(12): 144-147.

[6] Fang Shenghui, Le Yuan, Liang Qi et al. Chlorophyll inversion of mixed vegetation based on continuous wavelet analysis. Journal of Wuhan University (Information Science), 2015, 40(3): 296-302.

[7] Yong-qi Wang, Hui Zhang. The Research of Speech Recognition in Low SNR Based on GA-SVM [J]. Applied Mechanics and Materials, 2014, 50(9): 723-731.

[8] Chang J, Kim N S, Mitra S. V. Oice Activity Detection Based on Multiple Statistical Models [J]. Signal Processing, IEEE Transactions on, 2006, 54(6): 1965-1976.

[9] Lu Yuanyao, Zhou Ni, Xiao Ke, etc. Improved speech endpoint detection algorithm under strong noise environment [J]. Computer applications, 2014, 34(5): 1386-1390.

[10] P. K. Pal and S. Phadikar, "Modified energy based method for word endpoints detection of continuous speech signal in real world environment [C]." 2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Kolkata, India, 2015: 381-385.