INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING
DOI: 10.46300/9106.2022.16.42 Volume 16, 2022

Background speech synchronous recognition method of e-commerce platform based on Hidden Markov model

Pei Jiang, Dongchen Wang*
Anhui Technical College of Mechanical and Electrical Engineering
Wuhu 241000
China
*Corresponding author's email: 58131184@qq.com

Received: June 25, 2021. Revised: December 11, 2021. Accepted: January 11, 2022. Published: January 12, 2022.

Abstract—In order to improve the effect of e-commerce platform background speech synchronous recognition and solve the problem that traditional methods are vulnerable to sudden noise, resulting in poor recognition effect, this paper proposes a background speech synchronous recognition method based on Hidden Markov model. Combined with the principle of speech recognition, the speech feature is collected. Hidden Markov model is used to input and recognize high fidelity speech filter to ensure the effectiveness of signal processing results. Through the de-noising of e-commerce platform background voice, and the language signal cache and storage recognition, using vector graph buffer audio, through the Ethernet interface transplant related speech recognition sequence, thus realizing background speech synchronization, so as to realize the language recognition, improve the recognition accuracy. Finally, the experimental results show that the background speech synchronous recognition method based on Hidden Markov model is better than the traditional methods.

Keywords—Hidden Markov; e-commerce platform; speech synchronous recognition;

I. Introduction

The voice is not only a shortcut for people to exchange and connect information, but also a unique function of human beings. It is also a communication tool often used by human beings. With the advent of modern information age, the use of intelligent technology for voice storage, recognition and synthesis can make the voice information be effectively used [1]. The importance of speech greatly promotes the development of speech signal processing. Speech recognition as an important field of signal processing research, its role is to convert speech into control commands, so that the computer and human voice fusion, speech recognition is applied to many technical fields, and even can be extended to e-commerce, with the rapid development of computer technology, speech recognition has become a hot issue in the field of science and technology applications, and gradually into people's daily life, speech recognition has been successfully applied to mobile phones, televisions and other intelligent devices, which has a profound impact on human life style in the future [2]. With the rapid development of wireless network, the development of mobile e-commerce is faster and faster, but at the same time, there are many security problems, such as user information being eavesdropped, intercepted, tampered with, etc.; At the same time, due to the relevant laws and regulations are not perfect, the security problems of mobile e-commerce seriously restrict the development of mobile e-commerce [3]. Many mobile e-commerce platforms are intermingled, and there are many security risks, such as service and content authenticity being difficult to distinguish, businesses publishing false advertisements, lack of strict audit, which seriously affect the normal operation and management of mobile e-commerce enterprises. E-commerce speech synchronous recognition is to convert all speech data into text form, break through the barriers of communication between machines and people caused by different languages and tones, and make e-commerce speech interactive recognition method become an important tool for human-computer dialogue. Document [4] proposed an intelligent speech recognition method of different frequency segments based on LPCC. First, the linear prediction inverted spectrum coefficient is extracted by the linear prediction model, describe the voice characteristics in detail, then capture the frequency visual spectrum information of speech through the Meyer frequency inverted spectrum coefficient, finally select the appropriate acoustic basic elements, build the HMM acoustic model, and the acoustic model obtained is centrally tested through the MLE criterion, to give full play to the role of the trainability, scalability and accuracy of the model, so as to realize the speech recognition of different frequency segments. Document [5] proposed a high precision recognition of terminal fuzzy speech based on semantic association. According to the fuzzy speech recognition principle of the terminal, the voice

signal is processed from the perspective of time domain and frequency domain, and relevant data is stored using circular queue to ensure that large-capacity voice data is stored in the limited capacity data area. Then the voice signal is divided into short-time signals of one frame plus window, and the time series reorganization is performed to extract the signal characteristics. Discover the correlation between different semantic blocks, analyze the statement semantics, select syllable syllables, seminal syllables, phonemes and words as the motifs, and correct the wrong speech according to the specific process of fuzzy semantic speech recognition, thus realizing speech recognition.

Since e-commerce background speech characteristics are very different from human speech characteristics, the above speech recognition method is susceptible to sudden noise and has poor recognition effect. Therefore, this paper proposes a background speech synchronization recognition method for e-commerce platform based on hidden Markov model. Analyze the overall structure of voice feature acquisition, preprocess voice data through conversion, pre-emphasis, frame and window, endpoint detection, collect voice features, measure the similarity of voice data itself through correlation functions, match background voice of e-commerce platform; use a low-pass filter to prevent overlapping distortion, use hidden Markov model (HMM) to window the voice unit matching, reduce the impact of sudden noise, correctly convert received voice signal to text form, realize e-commerce background voice recognition. Experimental results show that this method can improve and improve speech recognition capability and can be widely used in various fields of speech recognition.

II. Background Speech Synchronous Recognition Method for E-commerce Platform

A. Feature extraction of speech tone

In order to ensure the effect of speech synchronous recognition under the background of e-commerce platform, it is necessary to collect the speech recognition features on the e-commerce platform [6]. If each frame of speech contains the redundant voice packet of the previous frame, when the packet loss problem is found in the recognition process, the redundant voice packet can be taken out and the lost voice packet can be recovered, but the redundant voice packet will aggravate the data transmission burden of the method. Therefore, a low bandwidth speech coding algorithm is needed to compress the previous frame speech, otherwise, the redundant speech packet is simply lost [7]. Based on this model, the network is running well at present, the packet loss is very small, and the voice frame does not need to transmit redundant voice packets. When the packet loss rate reaches a certain value, the method needs to add redundant voice packets to each frame of voice, so that the method can be used for packet loss recovery [8]. Therefore, the method transmits voice packets of variable length, and the method can distinguish voice packets by recognizing the length of each voice packet whether to add redundant voice packets or not and to recover the lost packets, we only need to consider the extended information of adding redundant voice packets in the voice frame design [9]. On this basis, the voice packet buffer information management mode under the background of e-commerce platform is optimized, as shown in Fig. 1

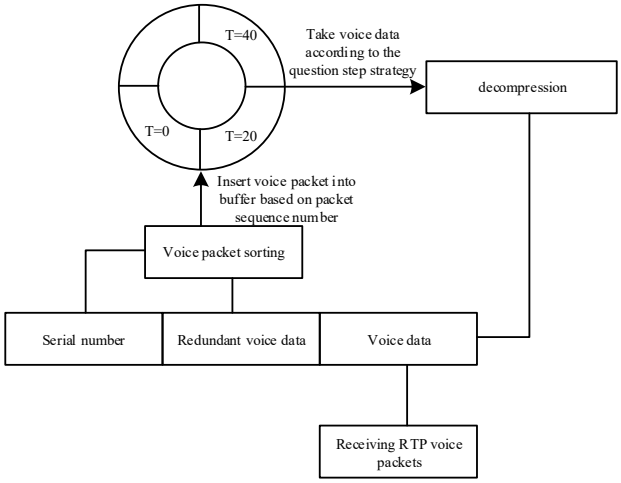

Figure 1. E-commerce platform background voice packet buffer management

When decompressing voice packets, voice packets are sorted according to the packet serial number before inserting jitter buffer. The original serial number of RTP packet is 16 bits, and the maximum value is only 65536. Taking random number as the initial value, it is easy to produce data overflow, which may lead to abnormal interruption of the program, or restart a new cycle [10]. Therefore, the extended sequence number is used as the sorting standard, which needs to be extended to 32 bits and then inserted into the jitter cache. In order to realize the real-time transmission of voice, we modify it to meet our own needs. Further, the network word order is designed as the sequential storage based on RISC chip, so that Intel PXA255 chip can correctly understand it [11]. Due to the space, the specific data structure of the real-time transmission module will not be listed here. Some key data structures are shown in Fig. 2.

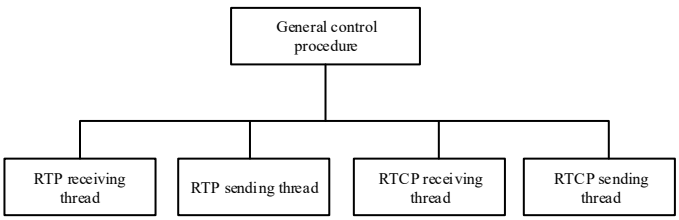

Figure 2. Overall structure of voice feature acquisition

In the process of speech recognition, the preprocessing of speech tone data aims to improve the quality of speech signal, unify the format of speech signal, and lay the foundation for subsequent speech signal feature extraction and intelligent recognition [12]. The pre-processing work of intelligent recognition of speech data features in multimedia network includes four steps: conversion of speech data, pre emphasis, framing and windowing, and endpoint detection. Voice data is essentially analog signal. Multimedia can only analyze digital signal [13]. Therefore, it is necessary to convert the voice data to convert analog signal to digital signal. The conversion process of voice data is shown in Fig. 3.

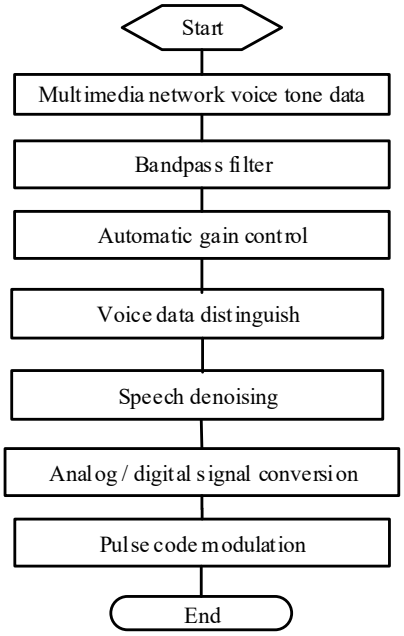

Figure 3. E-commerce background voice acquisition process

In order to facilitate the transmission and recording of signals, the method of deliberately comparing the amplitude of one frequency band with that of another frequency band to increase it is called pre emphasis. Among them, the first, second and third resonant peaks are the most obvious changes in tone [14]. The average, maximum and minimum of the three resonant peaks are selected as the characteristic parameters of multimedia network speech signal. The steps of extracting formant characteristic parameters are shown in Fig. 4.

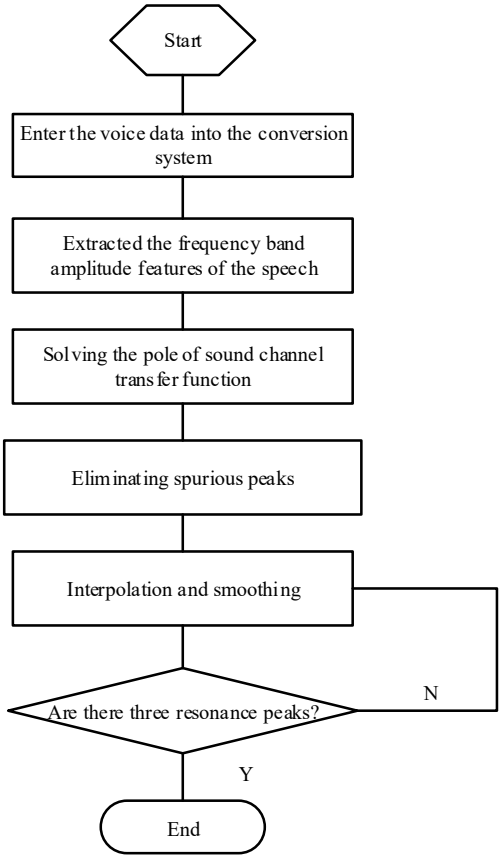

Figure 4. Extraction steps of speech synchronization feature parameters

As shown in Fig. 4, the voice data is input into the conversion system, extracting the frequency band amplitude characteristics of the voice through the linear prediction coefficient, obtain the mean, maximum and minimum of the resonance peak, eliminate the false peak, filter the voice data with interpolation and smoothing, judge whether the speech segment has three resonance peaks, if output the resonance peak characteristic parameters as the voice tone changes, if otherwise go back to the previous step to filter processing. Based on the feature extraction step of speech synchronization, the speech features are collected and analyzed, so as to improve the recognition accuracy. The cepstrum coefficient of hidden Markov model is a parameter which can reflect the characteristics of multimedia network speech tone data based on the human auditory frequency domain.

B. E-commerce platform background voice and matching

Further application of speech recognition technology in mobile e-commerce security often needs to build a comprehensive security model to ensure the security of mobile e-commerce. The speech signal is received, the noise in the speech signal is processed, and the speech features are extracted properly; the speech features previously stored in the database are compared with the extracted speech features [15]. The tone is one of the most important parts of speech. People's speech semantics will be different with different tones, and the emotions expressed will be different. Naturally, although people speak the same language, there will be obvious differences in tone. Therefore, it is not so much to extract the feature parameters of tone data as to extract the emotional feature parameters contained in speech take. This paper starts with the acoustic features, such as voice tone data feature extraction [16]. Acoustic characteristics include pitch frequency, resonance peak and hidden Markov cepstrum coefficient. In the field of speech recognition, the accurate extraction of gene frequency characteristic parameters is very important. Research shows that the tones produced by human voice contain different emotions, and the measure of emotion is gene frequency, so audio is one of the main characteristic parameters of voice tone data [17]. There are many methods to extract pitch frequency, but the autocorrelation function method is relatively the simplest and most efficient. Autocorrelation function, as the name suggests, is to measure the similarity of voice tone data itself through a correlation function.

$$f(v)=\sum_{m=0}^{N} t(r)t(r-2N) \tag{1}$$

In the formula, v is the delay of signal, $t(r)$ is a frame signal, and N is the window length of window function. When sound passes through a resonant cavity, it is filtered by the cavity, which redistributes the energy of different frequencies in the frequency domain. Part of the reason is that the resonant effect of the cavity is strengthened, and the other part is attenuated. Due to the uneven distribution of energy, the strong part is like a mountain, so it is called a formant. Formant is the spectrum envelope of multimedia network voice tone data, which is another important parameter eigenvalue in acoustic characteristics [18]. Under the effect of oronasal radiation and glottic excitation, the 800

Hz high frequency end of the average power spectrum of speech tone data will decline by 6 dB / times, which seriously affects the calculation of its high frequency end. Therefore, after the conversion of speech tone data, it is necessary to pre emphasize it to make up for the spectrum decline and make the spectrum flat. The z-transfer function of pre emphasis is as follows

$$f(z)=1-kz^{-1} \tag{2}$$

In the formula, K is the pre weighting coefficient. As a whole, the processed voice tone data has been in a state of fluctuation and lack of stability. If it is analyzed directly, it will directly affect the accuracy of recognition. However, if a certain segment (10 ms ~ 30 ms) is intercepted, the voice tone data state is stable [19]. This kind of stability is called temporary stability, so as long as the data in the temporary stable state is intercepted by segments in this way, we can use the processing method of stationary signal to process, which is framing. Generally speaking, segmented interception is realized by window function with limited length. The frame length of each segmented voice tone data is the same and sorted according to time series. The common window functions are $x(n)$

Among them, the rectangular voice window is:

$$x(n)=\begin{cases}0, & \text{other}\\ 1, & 0\le n\le N-1\end{cases} \tag{3}$$

Hidden Markov model is:

$$x(n)=\begin{cases}0, & \text{other}\\ 0.54-0.46\cos\left[\dfrac{2\pi n}{N-1}\right], & 0\le n\le N-1\end{cases} \tag{4}$$

Hanning window is:

$$x(n)=\begin{cases}0, & \text{other}\\ 0.5-0.5\cos\left[\dfrac{2\pi n}{N-1}\right], & 0\le n\le N-1\end{cases} \tag{5}$$

In the formula, N is the window length. Different window functions will have a direct impact on the analysis results. Based on this, the hidden Markov model is used for framing, because the hidden Markov model has less spectrum leakage. After preprocessing the tone data, it is necessary to extract the feature parameters. After receiving the voice information processed and extracted by the voice service provider, the mobile e-commerce enterprise can use relevant methods to analyze the voice data information, and then judge whether the user name exists or not [20, 21]. If the existence of the user name is removed, it means that the database has recorded and stored the user's voice feature information; by comparing and analyzing the latest voice feature information and new data before, if the two groups of voice features are consistent, the method is adopted to identify the user's identity, which indicates that the user's identity is legal. Let $f(x)$ be the window function, $t(x)$ be the frame signal, where x is the frame sequence, and $f(x)$ $t(x)$ be the voice frame signal after windowing. The background speech synchronous recognition method based on Hidden Markov model is to automatically select the window function form according to the user's speech characteristics. The part of speech decoding and grammar analysis are carried out under the hidden Markov model, from which the speech signal frequency can be obtained. Let y be the frame sequence transformed by hidden Markov model, which is as follows:

$$P(y)=x(n)\sum_{x=0}^{X-1}\mathrm{f}(x)\mathrm{t}(x)e^{-\frac{2x\pi y}{X}} \tag{6}$$

Set the energy of speech feature processing as e, and the semantic parsing result after HMM processing is as follows.

$$P'(y)=\sum_{x=0}^{X-1}E\cos\left[\frac{\mathrm{x}\pi(x+0.5)}{XP(y)}\right] \tag{7}$$

There are some failure data in the analytical results obtained by the formula, so it is necessary to delete some data, as shown in the formula.

$$P=\arg P'(y)\max P(\mathrm{a},b)\times\Pi P(\mathrm{a}_1,\mathrm{a}_2,\mathrm{a}_3) \tag{8}$$

Where p (a, b) is the result of normalization. Fuzzy speech file is an important dependent factor in context sensitive block transmission, which can clearly reflect the recognition differences of different spectrum sequences. In different contexts, the fuzzy speech signal emitted by the sound source may have a large deviation from the actual transmission intention, and these deviations always present a centralized distribution in the block spectrum feature structure [22]. Generally, the larger the deviation effect is, the more obvious the aggregation in the spectrum feature structure is. Due to the physical properties of context sensitive blocks, such as translatable and rotatable, it can better adapt to the transmission characteristics of voice source in the spatial environment. Taking the above and below sensitive blocks as the application environment, the hidden Markov filter combines the real and imaginary part vectors of the sound source to obtain the accurate transmission preconditions on the basis of fully adapting to the characteristic structure of the speech spectrum. Let q represent the real part and imaginary part vectors of the fuzzy speech source, and w 'represent the spectral characteristic coefficients of the context sensitive blocks

$$\Delta P=\frac{t^{r-1}\dfrac{Q_1}{Q_2}\cdot e}{\sqrt{y^2+q^2}} \tag{9}$$

Among them, e represents the original output frequency of the fuzzy speech source, t represents the recognition organization derivative in the block, r represents the fixed physical power coefficient, and y and q represent two different recognition basis biases. Assuming that in context sensitive block environment, the directional offset coefficient provided by the filter transmission condition is f, and the data condition within the basic sound source frame is f, the simultaneous formula χ can express the data amplitude and pitch vector of the fuzzy speech as follows:

$$\begin{cases}|\overline{u}|=\prod_{P=1}^{i\to\infty}f\dfrac{\sqrt{(R'-R'')}}{\Delta P\dot{e}}\\ |\overline{w}|=\dfrac{[a+(1-\chi)]'}{s_1^2+s_2^2}\end{cases} \tag{10}$$

Among them, u represents the data amplitude of fuzzy speech, w represents the pitch vector of fuzzy speech, i represents the maximum range of transmission conditions of hidden Markov filter, R represents the ideal speech input coefficient and real speech input coefficient, e represents the original transmission mean condition of speech data, s

represents the intra frame parameter data of two different sound sources, and a represents the stored speech output coefficient of sound source. Let the total amount of intra frame data of sound source is ϕ, when the speech recognition frequency is complete, the feature descriptor of fuzzy speech can be expressed as:

$$A=\int_{|\vec{u}|=1}^{\infty}\int_{|\vec{w}|}^{\infty}\frac{|\vec{u}|\sqrt{\lambda\cdot|\phi|}}{|\vec{w}|d^2-\beta a^2} \tag{11}$$

In the above formula, d represents the restriction coefficient of context sensitive block to fuzzy speech, β represents the precise recognition offset, and a represents the maximum speech recognition vector. There are two basic types of feature cues: the time difference between the ears of the sound source and the intensity difference between the ears of the sound source. The inter ear time difference of sound source refers to the directional difference of two fixed time nodes detected by the ear structure in the process of recognition and transmission of fuzzy speech, which is directly affected by many physical factors such as the propagation distance of sound source to the ear structure. In order to find the intensity difference of fuzzy speech, it is necessary to keep another speech feature clue variable —— time difference unchanged, so assume that the transmission state of fuzzy speech data remains unchanged in a context-sensitive region environment, that is, the intensity difference between sound sources is determined by the time difference between ears and the associated transmission medium coefficient. Let k represent the time node when the fuzzy speech arrives at two ear structures [23].

$$|k_1-k_2|=\frac{j}{l}[\sqrt{(g\cdot A)}]^{\mu} \tag{12}$$

Where, i and j represent two unequal fuzzy speech feature parameters, and g represents the transmission state condition of fuzzy speech data, μ represents the power term of time difference. On the basis of the formula, let x represent the transmission medium coefficient of the context sensitive area, and use x to express the calculation process of the characteristic clue of the intensity difference between the ears of the sound source as a formula.

$$z=\log\frac{\sum_{c=1}|k_1-k_2|^2}{\sum_{c=1}xb} \tag{13}$$

In the above formula, c represents the fixed identification parameter of ear structure, and b represents the specified direction vector of accurate identification definition.

C. Realization of background synchronous identification in e-commerce platform

When mobile e-commerce enterprises use speech recognition technology, they not only need to receive the read voice information, but also need to update the voice features in real time, seriously summarize new rules, and scientifically design relevant detection methods, so as to check the voice features of users regularly. Based on this, we need to compare and analyze these voice features, and find the small differences in time, so as to obtain more new voice features, and then put them into the database, and transmit them to the relevant enterprise database. If the method detects the voice information recorded by the same user, when the number of voice features exceeds 20, the corresponding formula needs to be used. If the conclusion is zero, the latest 20 information extracted again is processed repeatedly. The construction of background speech recognition method is completed on certain hardware conditions and experimental platform. Speech synchronous recognition is essentially a pattern recognition process, mainly including speech signal preprocessing. Its basic principle is shown in Fig. 5.

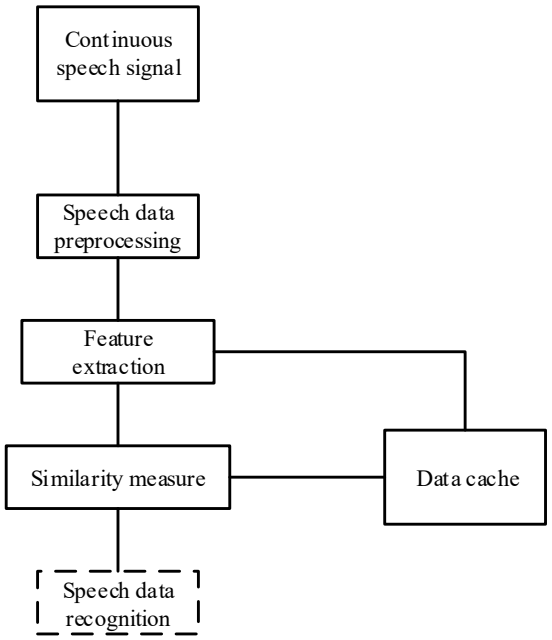

Figure 5. Basic principle of speech recognition

It can be seen from the figure that the background speech synchronous recognition method includes not only the core recognition program, but also speech input, parameter analysis and grammar language model construction. The speech recognition method is composed of speech signal preprocessing, core calculation and basic data recognition. The background speech synchronous recognition based on Hidden Markov model is to correctly convert the received speech signal into text form. The speech signal is time-varying and has stationarity. Therefore, when processing the speech signal, it is necessary to use function to process the background speech signal Each segment is called a frame, and there is a certain overlap between adjacent frames, which can reduce the jump. Extracting robust features of speech signal from each frame can complete noise elimination and feature extraction. The speech signal will change with time. Once the noise interference of aliasing distortion occurs, the speech signal processing will be invalid. Therefore, before synchronous recognition, it is necessary to use low-pass filter to prevent aliasing distortion. Voice signal is a non-stationary signal, affected by many factors such as glotres, sound channel and radiation. Traditional speech signal processing is based on linear system theory and the basic assumption is that the speech signal characteristics change over time and slowly over time. This assumption derives various "short-time" processing methods, and the speech signal is split into some short segments and reprocessed, each treated as a determined stationary signal, later yielding a new

time-dependent series used to describe the speech signal. As the research progressed, the speech signal was found to be a complex nonlinear process. With acoustic and aerodynamic theory analysis, voice not only has nonlinear vibration process, the tongue, the shape of the tongue, voice signal (especially friction sound, blasting sound, etc.) will produce vortex in the channel boundary layer, and eventually form turbulence, when other sound, valve airflow still has turbulence, and turbulence itself is a kind of chaos. The speech time-domain waveform has self-similarity and exhibits periodicity and randomness, which is also a manifestation of the fractal structure of the speech signal consistent with the human auditory neural signal transceiver. Therefore, to extract voice features, the influence of voice gate, channel and radiation on speech are handled through three steps: filtering, sample acquisition and construction of voice framework.

Hidden Markov model (HMM) is used to windowing the unit matching, which can make the signal transmission between adjacent frames smoother. The processed results are affected by the sudden noise, and the short-term average energy of some speech frames suddenly increases, which makes the recognition results inaccurate. Therefore, the processing flow as shown in Fig. 6 is designed.

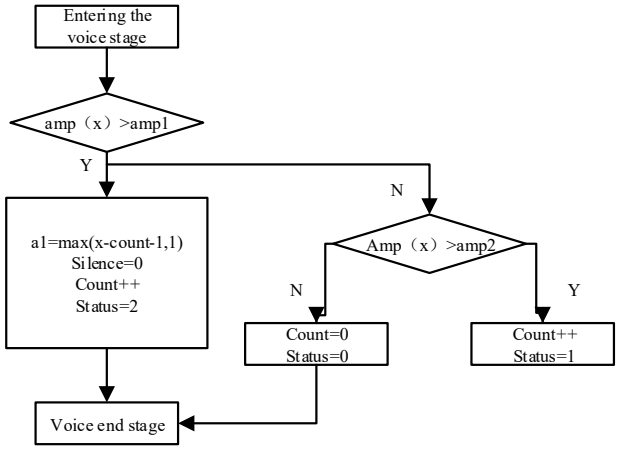

Figure 6. Speech recognition process

According to the design process of the method software, the hidden Markov model is used to windowing the unit matching, which can make the signal transmission between adjacent frames smoother. The window function shape is automatically selected to obtain the frame sequence transformed by hidden Markov model. Because there are some invalid data in the obtained results, some data needs to be deleted, and the processing flow is designed. Due to the open characteristics of wireless network, its application in mobile e-commerce will often cause a series of security risks to communication, resulting in security problems in the transmission of voice information, such as eavesdropping or tampering, which seriously affects the security of information. Therefore, it is necessary to adopt scientific and effective measures to improve the confidentiality, integrity and security of voice information. In order to improve the security of mobile e-commerce, information hiding technology can be used to enhance the security of voice information by using image or audio as hiding carrier and encryption. It is worth noting that if the picture is used, it is necessary to convert the format, convert the voice information into a readable format, mark and insert the end, beginning and content of the information appropriately, so as to avoid the change of the picture format. Finally, the encrypted information label format and pictures are sent to the decryption method of mobile e-commerce enterprises, and stored in the database after processing.

III. Analysis of Experimental Results

In order to analyze the effectiveness of the background speech synchronous recognition method based on the Hidden Markov model, we need to extract part of the speech training set from the standard pattern recognition database. The setting of experimental parameters is shown in the Table 1.

Table 1. Experimental parameter setting

Parameter		Numerical value
SAMpling frequency		25kHz
Voice signal window	Hidden Markov model	24 dimensions
	Length	20ms
	Framing	250 spots
	Frame shift	80 spots
Parametric vector quantization		50 code

In order to enhance the standardization of the experiment, the relevant experimental parameters are set according to the following Table 2.

Table 2. Experimental parameter setting table

Parameter name	Numerical value
Block environmental conditions	Context sensitive block
Output frequency band of sound source	650 MHz-800 MHz
Output wave number of sound source	0.78 μF
Experiment time	55 min
Speech cutting parameters	0.22
Maximum value of signal segmentation rate	72.3%-85.6%
Speech recognition coefficient	0.45
Maximum depth of sound source signal	8.10×10^{-7} μm

In order to make the experimental results have more practical significance, the experimental parameters of the experimental group and the control group are always consistent. In order to prevent the speech synchronous recognition method installed on the computer from being affected by the hardware performance, it is necessary to unify the high-end method performance on the computer for experimental verification and analysis. According to the

above experimental parameters and environment, the document [4] method and hidden Markov model method are compared and analyzed under the influence of sudden noise. The voice signal and short-term energy of the two methods are verified, and the results are as follows.

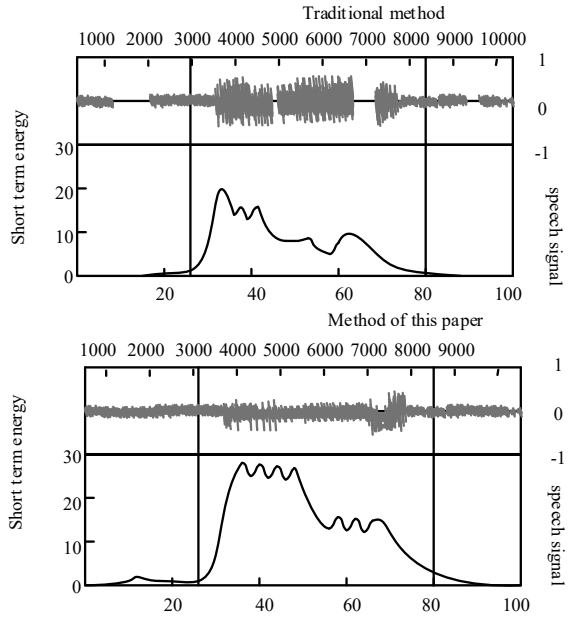

Figure 7. Detection results of two methods to identify terminal conditions

It can be seen from the Fig. 7 that: the document [4] method has the phenomenon of speech recognition interruption in the signal, which leads to short-term energy failure; while the hidden Markov model method has no interruption phenomenon, which can accurately obtain speech data. This is because this paper window the hidden Markov model for the nonlinear features of speech, smoothing the signal transmission between adjacent frames of the speech unit, reducing the sudden increase of the short-time average energy of some speech frames, so there is no interruption. Further, 60 min is taken as the experimental time, and the change of output speech recognition number is recorded in this period. According to the above comparison, the recognition effects of the two methods are compared under the influence of sudden noise, and the results are shown in the Table 3.

Table 3. Comparison of recognition effect of two methods

Noise/ dB	Document [4] method	Method based on Hidden Markov model
20	71%	86%
40	69%	89%
60	65%	91%
80	70%	92%
100	42%	90%

The comparison results show that the recognition effect based on HMM is above 86%, which is better than that of document [4] method, and the design of HMM Speech synchronous recognition method is effective, which fully meets the research requirements. This is because this paper first preprocesses speech data through conversion, pre-emphasis, frame and window, endpoint detection, collect speech features; uses low-pass filter to prevent voice distortion, uses hidden Markov model to window out the speech unit matching, reduce the impact of sudden noise, and together to improve the recognition effect.

IV. Concluding

In the current development process, mobile e-commerce also faces many security problems, especially management and technology problems, which seriously affect the security of information and data. Therefore, it is necessary to reasonably use speech recognition technology to effectively ensure the security of mobile e-commerce transactions through voice information storage and recognition, voice feature information transmission and update. In this paper, we propose the background speech synchronization recognition method of an e-commerce platform based on a hidden Markov model. Voice features are collected by preprocessing speech data; measure voice data similarity through relevant functions to match the background voice and tone of e-commerce platform; use hidden Markov model to window out the speech unit matching, reduce the impact of sudden noise, and realize the recognition of e-commerce background voice. The experimental results show that this method can effectively improve the speech recognition ability and has practical significance to ensure the security of mobile e-commerce transactions.

However, this paper is mainly limited to the research on improving the voice recognition effect of business platforms, and does not significantly improve the recognition speed. The future research can further accelerate the recognition speed on the basis of improving the recognition accuracy.

Acknowledgement

This work is supported by 2020 Anhui Provincial Quality Engineering Project -- Introduction to E-commerce (2020SZSFKC0259); 2019 Excellent Young Talents Support Program in Universities (GXYQZD2019107); 2020 Anhui Province "Grassroots Teaching Organizations" and "Basic Teaching Activities" standardized construction project -- Corporate Image Design (2021sfk10).

References

[1] K. G. Jahromi, D. Gharavian, H. Mahdiani, “A novel method for day-ahead solar powerprediction based on hidden Markov model and cosine similarity,” Soft Computing, vol. 24, no. 7, pp. 4991-5004, 2020.

[2] L. Liu, Y. Jiao, F. Meng, “Key Algorithm for Human motion recognition in virtual reality video sequences based on hidden Markov model,” IEEE Access, vol. 8, no. 10, pp. 159705-159717, 2020.

[3] Y. T. Tseng, S. Kawashima, S. Kobayashi, et al., “Forecasting the seasonal pollen index by using a hidden Markov model combining meteorological and biological factors,” The Science of the Total Environment, vol. 698, pp. 134246.1-134246.10, 2020.

[4] Y. Zhang, B. Li, X. Luo, et al., “Personalized mobile targeting with user engagement stages: Combining a

structural hidden Markov model and field experiment," Information Systems Research, vol. 30, no. 3, pp. 787-804, 2019.
[5] T. Chadza, K. G. Kyriakopoulos, S. Lambotharan, "Analysis of hidden Markov model learning algorithms for the detection and prediction of multi-stage network attacks," Future Generation Computer Systems, vol. 108, pp. 636-649, 2020.
[6] C. Djellali, M. Adda, "A new hybrid deep learning model based-recommender system using artificial neural network and hidden Markov model," Procedia Computer Science, vol. 175, no. 10, pp. 214-220, 2020.
[7] M. Xue, H. Yan, H. Zhang, et al., "Hidden-Markov-Model-Based asynchronous H-infinity tracking control of fuzzy Markov jump systems," IEEE Transactions on Fuzzy Systems, vol. 29, no. 5, pp. 1081-1092, 2021.
[8] Y. Lu, S. An, "Research on sports video detection technology motion 3D reconstruction based on hidden Markov model," Cluster Computing, vol. 23, no. 3, pp. 1899-1909, 2020.
[9] J. Tang, J. Hou, Y. Song, et al., "Effective exploitation of posterior information for attention-based speech recognition," IEEE Access, vol. 8, pp. 108988–108999, 2020.
[10] G. Sreeram, R. Sinha, "Exploration of end-to-end framework for code-switching speech recognition task: Challenges and enhancements," IEEE Access, vol. 8, pp. 68146-68157, 2020.
[11] L. Ma, "Construction of intelligent building sky-eye system based on multi-camera and speech recognition. International journal of speech technology," vol. 23, no. 1, pp. 23-30, 2020.
[12] A. Kumar, R. K. Aggarwal, "Discriminatively trained continuous Hindi speech recognition using integrated acoustic features and recurrent neural network language modeling," Journal of Intelligent Systems, vol. 30, no. 1, pp. 165-179, 2020.
[13] M. A. Khalighi, H. Akhouayri, S. Hranilovic, "Silicon-photomultiplier-based underwater wireless optical communication using pulse-amplitude modulation," IEEE Journal of Oceanic Engineering, vol. 45, no. 4, pp. 1611-1621, 2020.
[14] F. S. Cabral, H. Fukai, S. Tamura, "Feature extraction methods proposed for speech recognition are effective on road condition monitoring using smartphone inertial sensors," Sensors, vol. 19, no. 16, pp. 3481-3482, 2019.
[15] I. Yasin, V. Drga, F. Liu, et al., "Optimizing speech recognition using a computational model of human hearing: effect of noise type and efferent time constants," IEEE Access, vol. 8, pp. 56711-56719, 2020.
[16] G. T. Yadava, H. S. Jayanna, "Enhancements in automatic Kannada speech recognition system by background noise elimination and alternate acoustic modelling," International Journal of Speech Technology, vol. 23, no. 1, pp. 149-167, 2020.
[17] M. A. Tahir, H. Huang, A. Zeyer, et al., "Training of reduced-rank linear transformations for multi-layer polynomial acoustic features for speech recognition," Speech Communication, vol. 110, no. 10, pp. 56-63, 2019.
[18] N. Viswanathan, K. Kokkinakis, "Listening benefits in speech-in-speech recognition are altered under reverberant conditions," The Journal of the Acoustical Society of America, vol. 145, no. 5, pp. EL348, 2019.
[19] T. F. De Toledo, H. D. Lee, N. Spolaor, et al., "Web system prototype based on speech recognition to construct medical reports in Brazilian Portuguese," International Journal of Medical Informatics, vol. 121, pp. 39-52, 2019.
[20] Y. Y., Shi, J. Bai, P. Y. Xue, et al., "Fusion feature extraction based on auditory and energy for noise-robust speech recognition," IEEE Access, 2019, vol. 7, no. 10, pp. 81911-81922, 2019.
[21] M. M. Ismail, A. Alsayyari, "Performance analysis of optical CDMA wireless communication system based on double length modified prime code for security improvement," IET Communications, vol. 14, no. 7, pp. 1-9, 2020.
[22] V. Osadchyy, R. V. Skuratovskii, A. Williams, "Analysis of the mel scale features using classification of big data and speech signals," International Journal of Applied Mathematics, Computational Science and Systems Engineering, vol. 2, pp. 52-63, 2020.
[23] J. S. Jakati, S. S. Kuntoji, "A noise reduction method based on modified LMS algorithm of real-time speech signals," WSEAS Transactions on Systems and Control, vol. 16, pp. 162-170, 2021.