Using speech recognition to analyze the physical condition of the elderly and its application in telemedicine

Jianyu Wang[1, *], Lingjing Yu[2, a], Ming Huang[3, b], Runchuan Feng[4, c], Yiyang He[5, d]

[1]Hangzhou Yanhai Education, Hangzhou City, Zhejiang Province, China

[2]Beijing Jiaotong University, Beijing City, China,

[3]Aquinas High School, CA, USA

[4]Nankai University, Tianjin City, China

[5]China Jiliang University, Hangzhou City, Zhejiang Province, China

*Corresponding author: yljqwjy@163.com

[a]Yu_08051005@163.com, [b]1243114016@qq.com, [c]543119345@. qq. com, [d]496852857@qq.com

These authors have contributed equally to this work

Abstract

With the continuous development of science and technology and the increasing importance of the elderly to their own health, telemedicine has been favored by more and more people and has gradually become a research hotspot in the medical field. Aiming at the lack of accuracy of telemedicine, this article starts with the basic situation of intelligent speech recognition and explains its application methods and methods; it uses the extraction of Mel frequency cepstrum parameters (MFCC) and dynamic time warping (DTW) algorithm to pair. The simulation experiment results show that the method can accurately match the corresponding semantics in the reference speech database. This provides a certain reference for improving the application level of the telemedicine system.

Keywords

Telemedicine, intelligent speech recognition, Mel frequency cepstrum parameters, dynamic time warping.

1. Background

With the accelerated development of network connectivity, the popularity of smart phones and the constant change of insurance standards, more and more medical service providers are starting to use electronic communication devices to complete their work, and the telemedicine industry has unprecedented opportunities for development. More than 15 million Americans received some form of telemedicine in 2016, according to the American Telemedicine Association.

Meanwhile, Venture capital has contributed to the industry's growth. Since 2013, telemedicine startups worldwide have raised more than $1.2 billion, according to CB Insights. In 2016, telemedicine companies raised $362 million globally. From January to April 2017, there were 18 investments in telemedicine worldwide, with an investment amount of 43 million US dollars.

Telemedicine has broken regional restrictions, made up for the shortage of medical resources in remote areas, further improved and improved the medical service level in big cities, and greatly promoted the development of medical treatment and health care undertakings. The UK government defines tele-care as "a service that allows people, especially the elderly and vulnerable individuals, to live independently and safely at home". It includes personal and environmental sensors to make people safer and more independent in their homes. 24-hour monitoring to ensure that this information can be transmitted to users in a timely and effective manner in the event of an emergency. Most remote care reduces the risk of adverse events and helps speed up response times. Some remote care, such as safety validation and lifestyle monitoring, have preventive functions and can detect signs of deterioration in the user's condition at an early stage.

Telemedicine differs from telemedicine in that it uses human-centered science and technology to enable people to receive care in their own homes. There is no international consensus on the definition and use of telemedicine. In the UK, it is based on a framework of social care, while in other countries, telemedicine is used in telemedicine practices.

As for DURE, which is an artificial intelligence assistant released by Baidu. It is widely used in China because of its novelty and convenience. Its main function is to become a Bluetooth speaker, providing HD video calls, video playback, remote monitoring and so on. For example, DURE controls furniture in the home, directs directions in public places, and the Academy uses it to record vast amounts of data. It can connect multiple smart devices and cooperative devices under Baidu (smart speakers, smart tablets, Bluetooth vehicle-mounted devices, Bluetooth speakers and headphones, etc.). through the connection between APP and device, users can easily and fluently communicate with device or mobile phone in natural language, and get rich AI experience.

Statistics show that the size of China’s intelligent voice market exceeded 15 billion yuan in 2018. With the expansion of the intelligent voice application industry and the increase of market demand, it is expected that the size of China’s intelligent voice market will further accelerate its expansion in 2020. In terms of patent data, according to the analysis of related patents by the State Patent Office, by the end of 2019, there were a total of 22,119 speech recognition related patents, while the number of artificial intelligence related patent applications was 14,813, indicating that the technical output of intelligent speech is higher than the industry as a whole, and the development speed is faster than the industry as a whole.

Relevant departments further attach importance to the development of intelligent voice industry. In December 2018, the annual meeting of China Voice Industry Alliance was held in Shanghai. Representatives from the Information Software Department of the Ministry of Industry and Information Technology believe that intelligent voice is the earliest artificial intelligence technology. In the next step, the Ministry of Industry and Information Technology will further promote the development of core artificial intelligence technologies represented by intelligent voice, strengthen technological breakthroughs, promote the integrated application of the industry, optimize the development environment, and primitively promote the scale development of the intelligent voice industry.

In terms of standard construction, the INTERNATIONAL standard of full-duplex speech Interaction (ISO/IEC 24661 Information Technology-Full Duplex Speech Interaction) was officially approved by IBID., led by IBID., in collaboration with China Institute of Electronic Technology Standardization and Institute of Automation, Chinese Academy of Sciences, at the ISO/IEC JTC 1/SC 35 plenary session held in Pusan, Korea from January 13 to 17. This standard has also become the first international standard of intelligent voice interaction in the field of human-computer interaction led by China.

2. Method

2.1 Experimental data collection

2.1.1 Data acquisition

The sound data of this study are all collected by notebook computer, recording pen and other recording equipment in real life environment. Generally, the physical condition of the elderly is divided into two grades: good and poor. The voice samples of the elderly with different levels of physical condition are collected.

Due to some limitations, the sounds collected in this study were recordings that imitated the natural state of an elderly person reading the same text.

2.1.2 Data set construction

The analog speech signal is sampled with Sample frequency 48,000, and is discretized into digital speech sequence, which is stored in the device in wav. Format. In order to avoid the aliasing distortion in frequency domain, the period should be selected according to the bandwidth of analog speech signal (according to Nyquist sampling theorem). In the process of quantizing the discrete speech signal, it will bring some quantization noise and distortion. According to the physical condition of the elderly, the recorded audio is preliminarily labeled to construct a data set for subsequent processing. The purpose of this study is to discuss the voice recognition under the above two physical conditions.

The collected speech signal waveform is shown in figure 1,2,3,4. (figures 1 and 2 show the sound waves of men imitating healthy and weak tones respectively; figures 3 and 4 show the sound waves of women imitating healthy and weak tones respectively.).

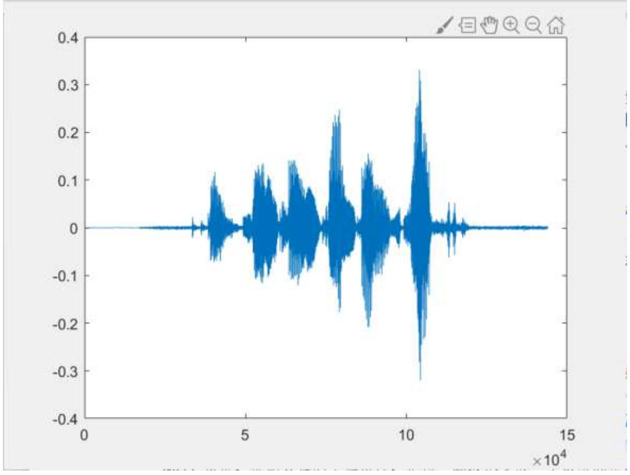

Fig. 1. the sound waves of men imitating healthy

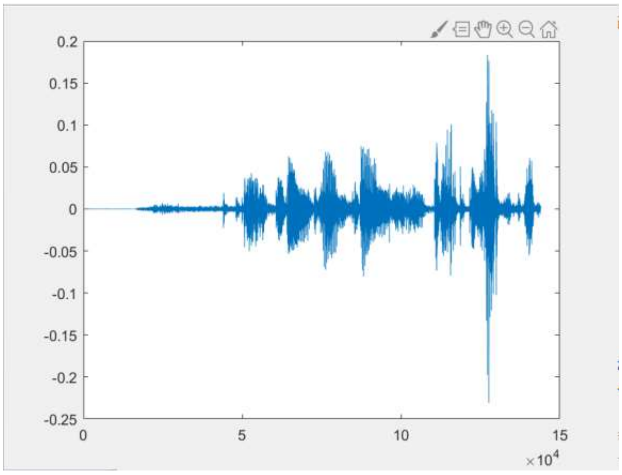

Fig. 2 the sound waves of weak tones respectively

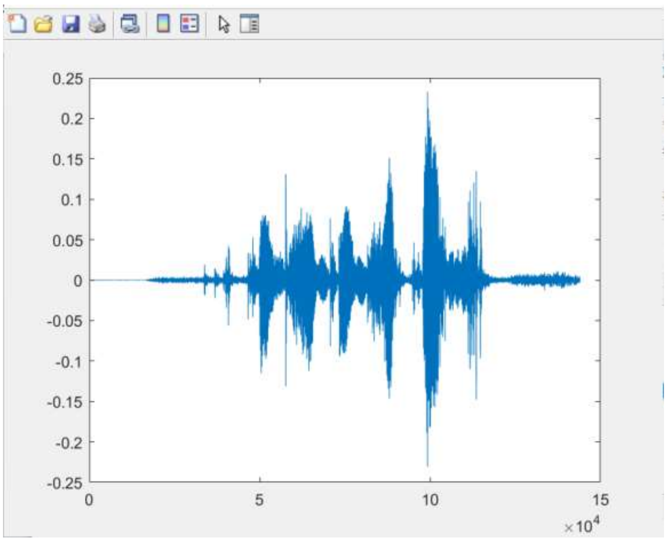

Fig. 3 the sound waves of women imitating healthy

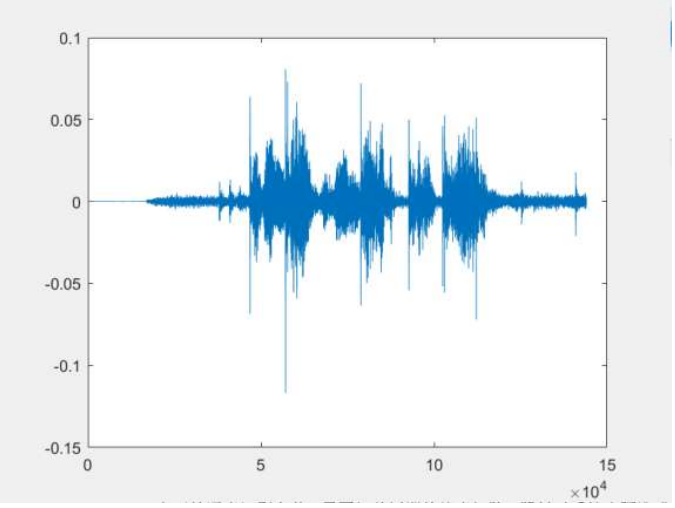

Fig. 4 the sound waves of weak tones respectively

2.2 Experimental data preprocessing

Before starting speech recognition, it is necessary to cut off the end of the silence to reduce the interference to the subsequent steps. To analyze the sound, we need to frame the sound, that is to cut the sound into small segments, each segment is called a frame. The framing operation is not a simple cut, but a moving window function. After framing, the voice becomes many small segments. However, the waveform has almost no description ability in time domain, so the waveform must be transformed. A common transformation method is to extract MFCC features and transform each frame waveform into a multi-dimensional vector according to the physiological characteristics of human ear.

2.2.1 Endpoint detection

Speech endpoint detection is to accurately find the starting point and end point of a speech signal from a speech signal, and distinguish the speech segment from the non speech segment. Since there are useless noise signals in the collected audio signals, endpoint detection is needed to improve the effect of speech analysis. In this study, a double threshold method is proposed for endpoint detection of speech signals.

2.2.2 Pre-emphasis

Because the average power spectrum of speech signal is affected by glottic excitation and oronasal radiation, the higher the frequency, the smaller the corresponding components. Therefore, it is necessary to improve the high frequency part of speech signal before analyzing it. The usual measure is to use digital filter to realize pre-emphasis.

In figure 5, a is the waveform of the speech, b is the curve after calculating the short-time energy of each frame of speech, and c is the curve after calculating the short-time zero-crossing rate of each frame of speech.

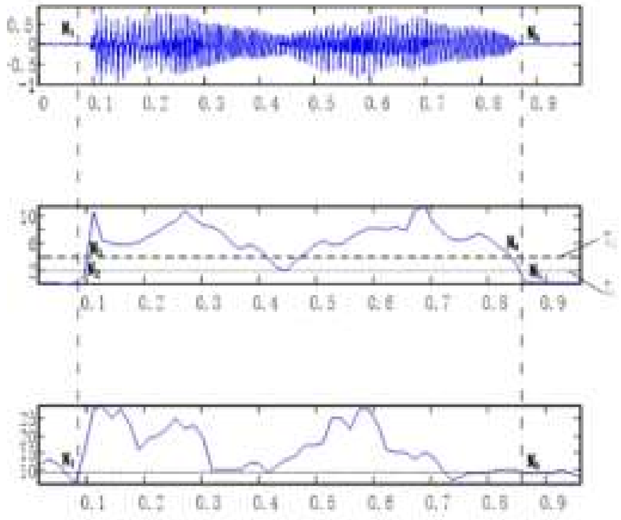

Fig. 5. Calculate the curve after the short-term zero-crossing rate of each frame of speech

2.2.3 Windowing and framing

The characteristics of speech signal change slowly with time, so the speech signal can be divided into some successive short segments (generally 10-30ms) for processing, that is, "short-term analysis". The speech signal is divided into segments to analyze its characteristic parameters, and each segment is called "a frame". The voice becomes a lot of little pieces. However, the waveform has almost no description ability in time domain, so the waveform must be transformed. A common transformation method is to extract MFCC features. According to the physiological characteristics of human ear, each frame waveform is transformed into a multi-dimensional vector, which can be simply understood as the vector contains the content information of this frame of speech.

The overlap between two frames is called frame overlap, and frame shift is the distance between the starting positions of the two frames. Usually, the limited length window is used to realize the framing. Several frames of speech correspond to a state. Every three states are combined into a phoneme, and several phonemes are combined into a word. In other words, as long as you know which state each frame of speech corresponds to, the result of speech recognition will come out.

2.3 Feature parameter extraction and analysis

After preprocessing the speech signal, it is necessary to analyze its characteristic parameters. In this study, Mel frequency cepstrum parameter (MFCC) is used to analyze the characteristic parameters of speech signal. Mel frequency cepstrum parameters (MFCC) calculation process: 1. Preprocessing pre-processing includes pre-emphasis, framing, windowing function. 2. After calculating the short-time. Fourier transform, the spectral line energy of the speech signal is obtained by taking the square of the frequency spectrum of the speech signal. The actual frequency scale is transformed into Mel frequency scale. The frequency response waveform of the Filter Bank is shown in figure 6. 3. The energy spectrum is passed through a group of triangular filter banks in mel scale, and the logarithmic energy of each filter bank is calculated. 4. MFCC coefficients can be obtained by discrete cosine transform (DCT). The MFCC coefficients per channel are shown in figures 7 and 8.

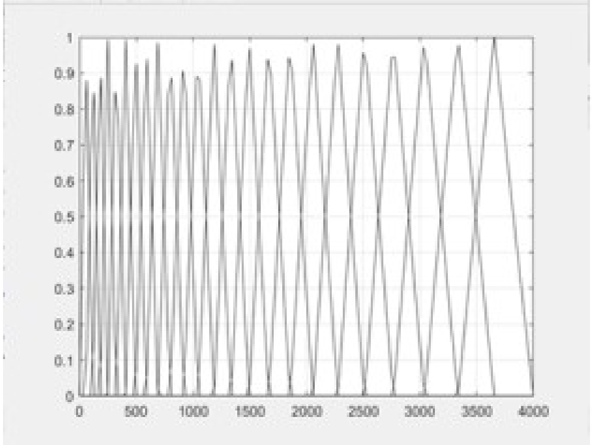

Fig. 6 Frequency response waveform of the filter ban

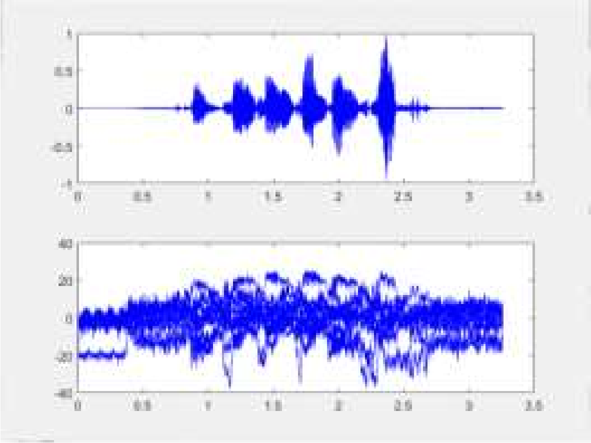

Fig. 7 good health

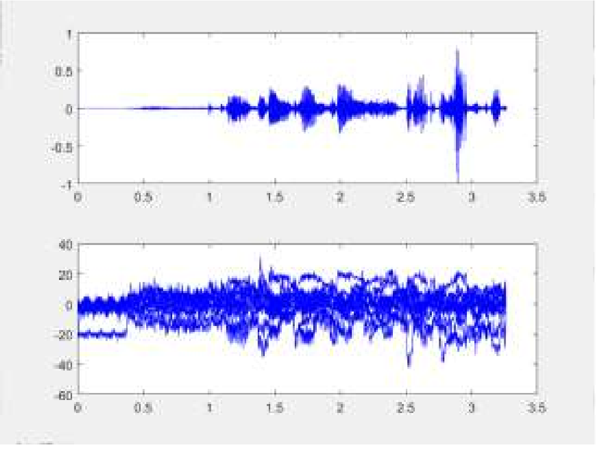

Fig. 8 poor health

2.4 Experimental Design

2.4.1 Database establishment

The speech signals collected in this study are divided into training set and test set, each part contains n pieces. After preprocessing and feature extraction, the speech signal is stored in the computer in the form of database, which constitutes the basic database of speech recognition system.

2.4.2 Pattern matching

Due to the change of speech rate, there is a nonlinear distortion between the output test speech and the reference mode, that is, compared with the reference mode, some phonemes of the input speech become longer while others are shorter, thus presenting random changes.

In template matching, the change of time length will affect the estimation of measurement, which will reduce the recognition rate. Therefore, effective time correction strategy is needed for time scaling. In this study, dynamic time warping (DTW) algorithm is proposed for matching.

DTW calculates the similarity between the preprocessed and framed speech signals and the reference template, and then calculates the distance between the two vectors according to a certain distance measure (generally Euclidean distance), so as to calculate the similarity between templates to find the optimal matching path.

Finally, Matlab programming simulation, to achieve four groups of different health conditions of the test speech recognition, can correctly match the corresponding reference to the reference voice database.

3. Conclusion

In this paper, a feature extraction algorithm based on spectrum analysis (MFCC) and a Dynamic time warping matching algorithm are proposed. it is reliable and feasible to use speech recognition to judge the health condition of old people.

4. Discussion

4.1 Current status of elderly care

With the aging and acceleration of the population, the aging process is intertwined with the contradictions of family miniaturization, empty nesting, and economic and social transformation.[1]The nursing problems of the elderly have become more prominent. The absolute amount of health care resources is inconsistent with uneven distribution, and the capacity and construction of the elderly medical and health service system are obviously insufficient. The number of geriatric hospitals, nursing homes for the elderly, rehabilitation hospitals, etc. are limited and unevenly distributed across regions.[1]At present, the level of medical services for the elderly in

China is still in a stage of development, which eventually caused their families to bear most of the pressure.

4.2 Current status and development of speech recognition

Speech recognition has been applied in many fields. As early as 2016, it has been proposed to automatically analyze and judge dietary conditions and food types through people eating and talking. Finally, a diet database was established, and a large number of experimental tests proved the feasibility of automatically classifying the types of food eaten and judging the diet while speaking, but there are also limitations such as people with insufficient speaking ability or people who cannot speak clearly. Will be considered poor diet [1]. Voice recognition is also developing rapidly in smart furniture and mobile phone assistants. In essence, it completes a series of follow-up operations by recognizing commands in its own vocabulary. At present, my country's intelligent speech recognition technology has entered a bottleneck period. The problems of noise interference, dialect recognition and fault tolerance in the environment cannot be solved well, which limits the development of speech recognition technology. [2] However, opportunities and challenges always coexist. Speech recognition faces various problems. Manufacturers and research institutes at home and abroad have invested a lot of money in research. It is believed that intelligent speech recognition technology will gradually progress with the development of science and technology, and finally be applied to All aspects of life.

4.3 Play a role in telemedicine for the elderly through voice Recognition

Officially, because of the low level of medical services for the elderly, voice recognition is used to make a judgment on the physical condition of the elderly to assist telemedicine. On the one hand, it can ensure the timeliness and on the other hand, it also greatly reduces the pressure on the family. There is currently a major problem in telemedicine: the combination of multimedia technology and medical equipment is not mature enough, and some simple telemedicine equipment has low detection accuracy and single test items, which cannot meet the multi-item testing requirements and standards required by telemedicine services. [1] If speech recognition can accurately and quickly analyze the health level of the elderly, it will be a great improvement for telemedicine.

References

[1] Yuchen Zhou, Yu Zhang. Article: Analysis onthe status quo and countermeasures of healthy old-age care in China[J]. Chronic Diseases Prevention Review,2019,3(11).

[2] Hantke Simone,Weninger Felix,Kurle Richard,Ringeval Fabien,Batliner Anton,Mousa Amr El-Desoky,Schuller Björn. I Hear You Eat and Speak: Automatic Recognition of Eating Condition and Food Type, Use-Cases, and Impact on ASR Performance.[J]. PloS one,2016,11(5).

[3] Hao Ouya, Wu Xuan, Liu Rongkai. The development status and application prospects of intelligent speech recognition technology [J]. Electroacoustic Technology, 2020, 44(03): 24-26.

[4] Zhang Xin, Wang Xiaohua. Research on the problems and countermeasures of telemedicine services under the background of "Internet +"[J]. Satellite TV and Broadband Multimedia, 2019(15): 34-35.

[5] Cang Yan, Luo Shunyuan, Qiao Yulong. Pig voice classification based on deep neural network[J].Journal of Agricultural Engineering,2020, 36(9): 195-204.

[6] Chen Jie. Speech feature parameters based on wavelet analysis and their application in speaker recognition[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2009.

[7] Liu Junfei. Research on Recognition Method of Depression Based on Speech Signal[D].t Tianjin: Tianjin Normal University, 2016.