Research on voice control and analysis technology for the elderly at home

Xiyang Chen*

Optoelectronic Information Science and Engineering
South China Agriculture University
Guangdong, China

*Corresponding author: 13710317000@163.com

Abstract. The current situation of our country is a big country with serious population ageing, and the proportion of the elderly living alone is the largest among them, most of the children of the elderly because of work needs, can not be taken care of around. Therefore, to improve the life quality of the elderly living alone, this study designed a voice-intelligent home system, using voice control technology and combining it with a microcontroller and Internet, which facilitates the family members of the elderly living alone to monitor the health condition of the elderly through this system and also provides convenience for the elderly to use the furniture.

Keywords: VAD endpoint detection, MFCC analysis, DTW principle, microcomputer

1 Research background

1.1 Population ageing status[1]

Since our country began to enter the population ageing 20 years ago, the number of the elderly population has become more and more huge, the speed of aging has become faster and faster, People's life expectancy has become higher and higher, the social population has shown the trend of aging. In May, the 2021 National Bureau of Statistics of the People's Republic of China released data from our country's seventh population census. According to the data, compared with 2010, the proportion of the population over 60 years of age in the total population increased by 5.44 percentage points, over 65 increased by 4.63 percentage points.

1.2 Living conditions of the elderly living alone[2]

The number of the elderly population in our country is huge, and most of them live alone. Their main source of income is their own retirement pension, in addition to which there are timely subsidies for their children, and their economy is relatively

R. Appleby et al. (eds.), *Proceedings of the 2nd International Conference on Intelligent Design and Innovative Technology (ICIDIT 2023)*, Atlantis Highlights in Intelligent Systems 10,
https://doi.org/10.2991/978-94-6463-266-8_17

independent, but their pension can only meet their living expenses, although the city has a good level of medical care, the elderly have medical insurance, but if there is no relatives in the side, and the city living interpersonal relationships are estranged, children can not take care of the elderly in an all-round way of life, and our country is not perfect in the development of old-age care institutions, and most of the elderly do not want to go to the old-age care institutions, they prefer to choose home care.

1.3 Research significance

In our country, there are a large number of elderly living alone. Due to their slow actions and responses, it is difficult for them to protect themselves quickly when encountering an emergency, and they are often liable to cause irreversible damage and may be in serious danger of their lives, the study of this paper can combine the living habits of the elderly living alone and their living conditions, design a voice-controlled intelligent home system, can help the elderly living alone life, and family members to monitor their health so that they can be quickly detected by voice control in case of danger.

2 System functions

The system is used for the elderly living alone, so the voice signal control is a more appropriate command input method. The specific flow of the system is shown in Figure 1, the system has two functions, one is through voice control technology and MCU driver operation instructions to facilitate the elderly to use the home, the other is through voice recognition technology to determine the health of the elderly, and through the internet to facilitate family monitoring of the health of the elderly. Corresponding to the two functions of the single-chip microcomputer, one is if compared for voice instructions, the single-chip microcomputer will input voice into the single-chip microcomputer driver, the single-chip microcomputer driver connection control circuit to control the home, such as lights and valves. The other is that if you compare the health status of the elderly, the microcomputer will analyze the health status through serial WIFI and routers to connect to the Internet, and then through the mobile server network sent to the family phone.

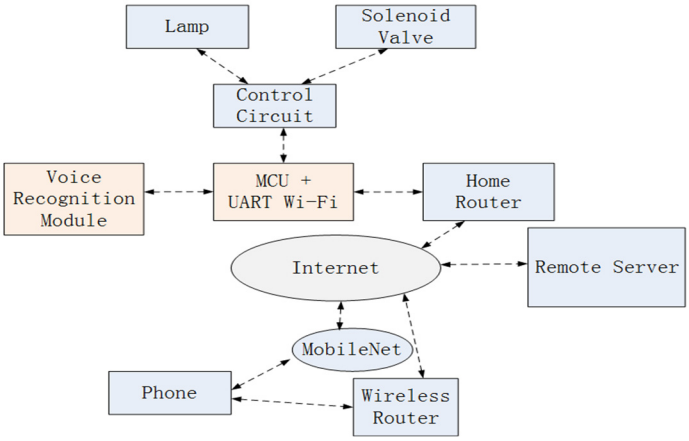

Fig. 1. Flow chart of system function

2.1 Voice-controlled smart furniture

The first function of the system is to use voice control technology, so that the machine can directly recognize the human voice through the command, and execute, where the need to use a single-chip computer to analyze the voice signal. But because machines, unlike human brains, can automatically filter out useless information and get useful instructions, the system here needs to use speech extraction and speech recognition methods in the speech recognition module, the obtained data are compared with the database, and the required instructions are obtained. The required instructions are input into the control circuit to realize the control of the intelligent home.

2.2 Judge Health status by sound

The second function of this system is to use the speech recognition technology, in the speech recognition module through the speech recognition algorithm, to identify the old voice behind the health status, these also through the SCM database comparison to determine the corresponding state of health of the elderly, and then through the SCM and serial WIFI connection to achieve the home wireless router connection, the router can connect with the Internet, so that the health of the elderly home can be sent to the family's mobile phone through the Internet, the health of the elderly family to monitor.

3 Principles of system

Voice control system through the Internet to remind the elderly physical status of the elderly and to facilitate the elderly to use the furniture function, this is mainly about the principle of speech recognition module. As shown in Figure 2, when the elderly voice signal, the voice input system, the system will extract the voice signal, first through the VAD endpoint detection, then the effective part is extracted by analyzing pitch period

with MFCC. Then the speech recognition based on DTW algorithm is the most important part of this research, which is the key to input into the MCU database for data analysis and comparison. At this point, the voice signal contains voice commands and the health of the elderly, and then the voice identified by the voice input into the MCU for analysis, the MCU compared the voice to get the language that the computer can recognize, finally, the function of controlling the home and informing the family about the health status of the elderly can be realized.

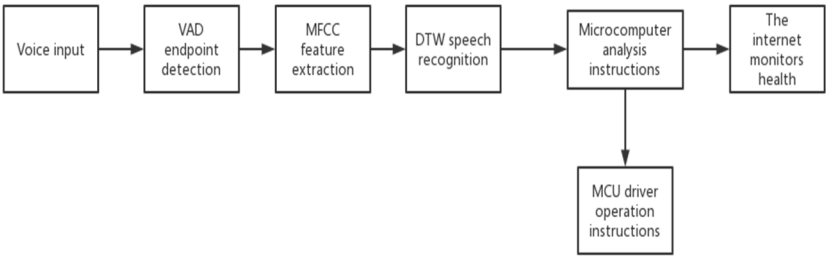

Fig. 2. System core schematic diagram

3.1 VAD endpoint detection[3]

In this paper, the whole speech endpoint detection is divided into 4 segments by double threshold method: mute, filter, speech segment and end. In the silent segment, if the short-term energy or zero-crossing rate exceeds the low threshold, start marking the starting point and enter the transition segment. In the transition segment, because the values of the two parameters are small, it is not certain whether they are in the real speech segment, so as long as the values of the two parameters fall below the low threshold, the current state will be restored to the mute state, if the high threshold is exceeded in both parameters of the transition segment, the speech segment is determined to be entered, i. e. the speech start and end points are detected. The process of VAD endpoint detection is shown in Figure 3. The system first analyzes the short-time energy and short-time average zero-crossing rate of the input voice signal, and selects a high energy threshold T2, then analyzes the short-time average zero-crossing rate of the input voice signal, most of the energy envelope of the speech signal is above this threshold, and the starting and ending points of the speech are outside the time interval corresponding to the intersection of the threshold and the short-term energy envelope, then a lower threshold T1 is determined according to the energy of background noise, and the two points where the energy curve intersects the threshold T1 for the first time are found respectively, find two points where the short-term average zero-crossing rate is below a threshold T3, which is the starting and ending point of a speech segment.

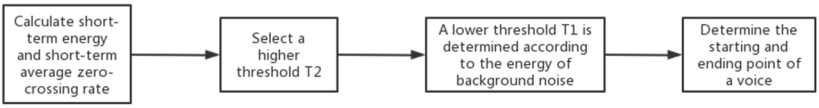

Fig. 3. VAD endpoint detection process

3.2 MFCC analysis

The output energy of each band-pass filter is taken as the basic feature of the signal, which can be further processed as the input feature of speech. The band-pass filter used in the MFCC analysis is the MEL filter banks. The MFCC analysis focuses on the auditory characteristics of the human ear, because the pitch of the sound heard by the human ear is not linearly proportional to the frequency of the sound, the Mel frequency scale is more suitable for human hearing. [6] the value of the so-called Mel frequency scale corresponds to the logarithmic distribution of the actual frequency. The detailed process of MFCC analysis is shown in Figure 4. The voice signal detected by the front end is filtered by Hamming window, and then the signal is input into the system, which Fast Fourier transform it, the transformed signal is then input into the MEL frequency filter bank for filtering, and the resulting signal is discrete cosine transform to obtain the cepstrum coefficient.[4]

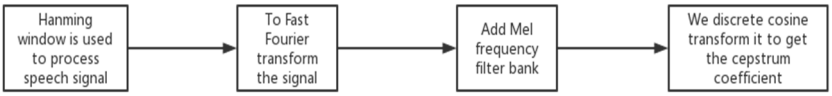

Fig. 4. MFCC analysis process

3.3 DTW principle

DTW is an earlier method of pattern matching. Based on the idea of dynamic programming, it solves the problem of template matching with different length when comparing speech signal feature parameter sequences in isolated word speech recognition. By calculating the similarity between the preprocessed and frame-divided speech signal and the reference template, then, according to some distance measure (usually European distance), the distance between two vectors is calculated to find the best matching path. So the regularization function with the minimum cumulative distance is obtained by DTW algorithm, which ensures the acoustic similarity between them. [7] the system will input the previously obtained cepstrum coefficients into the DTW algorithm for comparison with the database. The DTW algorithm comparison flow is shown in Figure 5. First, the distance matrix between the points of the two sequences is calculated, then look for a path from the upper left corner of the matrix (1,1) to the lower right corner (i, J) , so the starting condition is Lmin (1,1) = m (1,1) ,

and the recurrence rule is Lmin (i, j) = min { Lmin (i, J-1) , Lmin (I-1, J-1)} + m (I, J) , finally, the element and the minimal path are obtained, and the result of the comparison between the speech signal and the database is obtained to realize the speech recognition.[5]

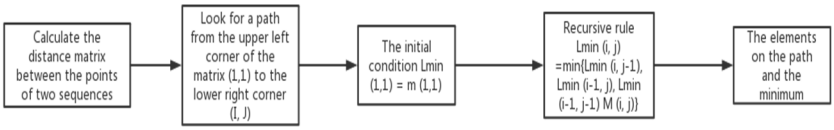

Fig. 5. DTW algorithm flow

4 Verification process

This study through the MATLAB R2020B DTW algorithm for training and testing, to verify the feasibility of speech recognition system. The data used in this paper are from computer voice card recordings of voice commands that the elderly living alone may use. The test consists of 80 speech sounds, divided into 8 groups, corresponding to 11 reference templates, each group of 10 speech sounds by different examiners (all of the examiners are female), to improve the universality of the test, and finally record the test results of each group and calculate the accuracy of recognition.

5 Data analysis

5.1 Generate Voice 1. wav to 10. wav, identification results as shown in Table 1

Table 1. Repeat “lights on” 10 times

Test the voice file	Identify the results	The identification result corresponds to the reference	Identify whether the result is right or wrong
1.wav	lights on	r_1.wav	Right
2.wav	lights on	r_1.wav	Right
3.wav	lights on	r_1.wav	Right
4.wav	lights on	r_1.wav	Right
5.wav	lights on	r_1.wav	Right
6.wav	lights on	r_1.wav	Right
7.wav	lights on	r_2.wav	Wrong
8.wav	lights on	r_1.wav	Right
9.wav	lights on	r_1.wav	Right
10.wav	lights on	r_1.wav	Right
The recognition accuracy is 90%			

5.2 Generate voice 11. wav to 20. wav, identification results as shown in Table 2

Table 2. Repeat "lights off" 10 times

Test the voice file	Identify the results	The identification result corresponds to the reference	Identify whether the result is right or wrong
11.wav	lights off	r_2.wav	Right
12.wav	lights off	r_2.wav	Right
13.wav	lights off	r_2.wav	Right
14.wav	lights on	r_1.wav	Wrong
15.wav	lights off	r_2.wav	Right
16.wav	lights off	r_2.wav	Right
17.wav	lights off	r_2.wav	Right
18.wav	lights off	r_2.wav	Right
19.wav	lights off	r_2.wav	Right
20.wav	lights off	r_2.wav	Right
The recognition accuracy is 90%			

5.3 Generate voice 21. wav to 30. wav, identification results as shown in Table 3

Table 3. Repeat "stop" 10 times

Test the voice file	Identify the results	The identification result corresponds to the reference	Identify whether the result is right or wrong
21.wav	stop	r_3.wav	Right
22.wav	stop	r_3.wav	Right
23.wav	stop	r_3.wav	Right
24.wav	stop	r_3.wav	Right
25.wav	up	r_4.wav	Wrong
26.wav	stop	r_3.wav	Right
27.wav	stop	r_3.wav	Right
28.wav	stop	r_3.wav	Right
29.wav	up	r_4.wav	Wrong
30.wav	stop	r_3.wav	Right
The recognition accuracy is 80%			

5.4 Generate voice 31. wav to 40. wav, identification results as shown in Table 4

Table 4. Repeat "hello" 10 times

Test the voice file	Identify the results	The identification result corresponds to the reference	Identify whether the result is right or wrong
31.wav	hello	r_5.wav	Right
32.wav	hello	r_5.wav	Wrong
33.wav	hello	r_5.wav	Right
34.wav	hello	r_5.wav	Right
35.wav	turn right	r_6.wav	Wrong
36.wav	hello	r_5.wav	Right
37.wav	hello	r_5.wav	Right
38.wav	hello	r_5.wav	Right
39.wav	hello	r_5.wav	Right
40.wav	turn right	r_6.wav	Wrong
The recognition accuracy is 80%			

5.5 Generate voice 31. wav to 40. wav, identification results as shown in Table 5

Table 5. Repeat "no" 10 times

Test the voice file	Identify the results	The identification result corresponds to the reference	Identify whether the result is right or wrong
41.wav	no	r_7.wav	Right
42.wav	no	r_7.wav	Right
43.wav	no	r_7.wav	Right
44.wav	no	r_7.wav	Right
45.wav	no	r_7.wav	Right
46.wav	turn right	r_6.wav	Wrong
47.wav	no	r_7.wav	Right
48.wav	no	r_7.wav	Right
49.wav	no	r_7.wav	Right
50.wav	no	r_7.wav	Right
The recognition accuracy is 90%			

5.6 Generate voice 31. wav to 40. wav, identification results as shown in Table 6

Table 6. Repeat "left" 10 times

Test the voice file	dentify the results	The identification result corresponds to the reference	Identify whether the result is right or wrong
51.wav	left	r_8.wav	Right
52.wav	left	r_8.wav	Right
53.wav	left	r_8.wav	Right
54.wav	left	r_8.wav	Right
55.wav	left	r_8.wav	Right
56.wav	left	r_8.wav	Right
57.wav	left	r_8.wav	Right
58.wav	yes	r_9.wav	Wrong
59.wav	left	r_8.wav	Right
60.wav	left	r_8.wav	Right
The recognition accuracy is 90%			

5.7 Generate voice 31. wav to 40. wav, identification results as shown in Table 7

Table 7. Repeat "turn right" 10 times

Test the voice file	Identify the results	The identification result corresponds to the reference	Identify whether the result is right or wrong
61.wav	turn right	r_6.wav	Right
62.wav	no	r_7.wav	Wrong
63.wav	turn right	r_6.wav	Right
64.wav	turn right	r_6.wav	Right
65.wav	turn right	r_6.wav	Right
66.wav	no	r_7.wav	Wrong
67.wav	turn right	r_6.wav	Right
68.wav	turn right	r_6.wav	Right
69.wav	no	r_7.wav	Wrong
70.wav	turn right	r_6.wav	Right
The recognition accuracy is 70%			

5.8 Generate voice 31. wav to 40. wav, identification results as shown in Table 8

Table 8. Repeat "down" 10 times

Test the voice file	Identify the results	The identification result corresponds to the reference	Identify whether the result is right or wrong
71.wav	down	r_10.wav	Right
72.wav	down	r_10.wav	Right
73.wav	down	r_10.wav	Right
74.wav	monkey	r_11.wav	Wrong
75.wav	down	r_10.wav	Right
76.wav	down	r_10.wav	Right
77.wav	down	r_10.wav	Right
78.wav	monkey	r_11.wav	Wrong
79.wav	down	r_10.wav	Right
80.wav	down	r_10.wav	Right
The recognition accuracy is 80%			

5.9 Comprehensive analysis

According to the analysis of the above speech recognition examples, when the sample data is 8 groups, which contains 10 tests of Chinese and English speech, the average accuracy of DTW speech recognition algorithm is 83.75%, each example has typical features, and the accuracy of each example is shown in Figure 6, which indicates that when the test samples are small and the content of speech signal is simple, each speech signal through DTW algorithm and the database the speech matching accuracy is about 80%, which is a relatively high accuracy rate and can prove the feasibility of the system for speech recognition.

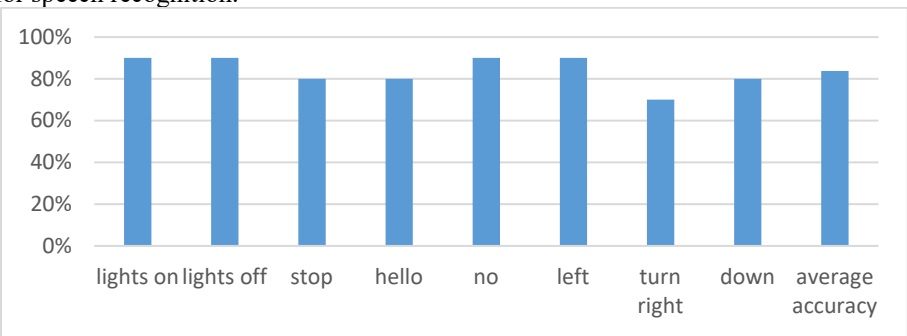

Fig. 6. The accuracy of each example

6 Conclusion

This paper describes an intelligent home system for the elderly living alone, which is based on the principle of voice control, it includes the functional design of the system and several algorithms of speech recognition, VAD endpoint detection, MFCC analysis and DTW principle. At the same time, using MATLAB R2020B speech signal DTW algorithm for simulation analysis, verify the feasibility. The simulation results show that the speech recognition accuracy reaches 83.75%, which indicates the feasibility of this study. However, the deficiency of this study is that when the speech content is gradually complex and similar, the recognition accuracy of this algorithm will be reduced, and this study only to the speech recognition part, additional analyses are needed from further experiments.

References

1. Yu, X.P., & Lu, B.Y. (2023). Analysis of Chinese population ageing trends and positive population ageing measures. Economic Forum (2),2.
2. Wong, Y.L., & Yang, C. (2021). A study on the composition and living conditions of the old people living alone in our country. Population and society, 37(05),12.
3. Nie, Q., Zhou, M.Q,. & Li, J.C. (2021). Improvement of double threshold speech endpoint detection algorithm in source localization. Electronic testing (15), 47-50. doi:10.16520/j.cnki.1000-8519.2021.15.015.
4. Li, G.J., Luo, P.Z., Qian, P., G., W.M., &Xing, M. (2023). Mel frequency cepstrum coefficient smoothing earphone equalization. Applied Acoustics,42(01), 67-75.
5. Kang E.M., & Mao K.N. (2023). Continuous map generalization method of linear feature based on DTW algorithm. Bulletin of surveying and mapping (04), 172-176. doi:10.13474/j.cnki.11-2246.2023.0125.
6. Juvela, L., Bollepalli, B., Wang, X., Kameoka, H., Airaksinen, M., Yamagishi, J., & Alku, P. (2018, April). Speech waveform synthesis from MFCC sequences with generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5679-5683). IEEE.
7. Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., ... & Keogh, E. (2012, August). Searching and mining trillions of time series subsequences under dynamic time warping. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 262-270).