Design of Multilevel Speech Automatic Recognition and Translation Software Based on Internet

Xin Xiong[1], Qing Xu[1]*, Mingjing Guo[2]

Xin Xiong: xiongx88@126.com, Qing Xu: xuq235@126.com
Mingjing Guo: guomj66@126.com

[1]Naval University of Engineer, WuHan,Hu Bei ,430033, China
[2]School of Science,East China University of Technology,Nan chang, Jiang Xi,China

Abstract: Speech recognition is an important research field that involves converting spoken language into textual form for computers to understand and process. In recent years, with the continuous development of embedded technology, embedded speech recognition technology has become the main research direction in the field of speech recognition. In this context, this article designs and implements an English translator speech recognition system. Our experimental results indicate that our designed translator performs well in terms of average speech recognition accuracy, reaching approximately 91.24%. In contrast, the average speech recognition accuracy of the traditional method 1 translation system is about 76.73%, while the average speech recognition accuracy of the traditional method 2 translation system is about 65.34%. These results indicate that the translator designed in this article performs better in speech recognition and has stronger speech recognition capabilities. This is crucial for the entire translation process, as more accurate speech recognition is expected to lead to higher quality translations, thereby improving translation effectiveness.

Keywords: Internet; English; MFCC; Translator; speech recognition system

1 Introduction

With the widespread popularity of the Internet and the continuous increase in global communication, automatic speech recognition and translation technology has become an important research direction in the field of information processing. The development of this technology not only eliminates language barriers and promotes communication between different countries and cultures, but also brings enormous convenience in fields such as commerce, education, healthcare, and entertainment. In this context, we propose the design of a multi-level speech automatic recognition and translation software based on the Internet. This system aims to combine advanced speech recognition technology with a powerful translation engine to provide users with a comprehensive and efficient language translation solution. By converting speech into text and then translating the text into the desired target language for users, the software has the potential to achieve seamless cross language communication[1]. This article will explore in detail the design, implementation, and performance evaluation of the system, with a focus on its multi-level structure and internet integration function. We will discuss the main components of the system, including voice acquisition, speech recognition, text translation, and user interface. In addition, we will also evaluate the performance of the

NMDME 2023, October 13-15, Xi'an, People's Republic of China

DOI 10.4108/eai.13-10-2023.2341325

system, especially in terms of speech recognition and translation accuracy. Through this study, we aim to provide users with a convenient voice translation tool, broaden their language communication skills, promote cross-cultural communication, and meet the needs of different fields. At the same time, this will also promote the further development of speech automatic recognition and translation technology, opening up new possibilities for future internet applications[2].

2 Technical basis for speech recognition system

Modern education in China is developing towards informatization, and the creation of an information platform in English teaching is also a demand for modern teachers and students to innovate traditional English classroom models. Speech recognition plays an important role in computer translation software. The process of speech recognition is shown in Figure 1.

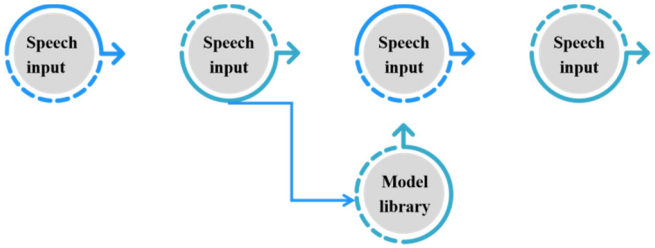

Figure. 1 The process of speech recognition

This can help students quickly understand the connotation of English knowledge, mainly including feature extraction, model training, and pattern matching, which is also a key consideration in the research process of this article. The general speech is processed, indexed, and transmitted by the system's auxiliary functions. There are significant differences between computers and natural languages[3]. How to accurately identify the differences between these two languages is a key problem that translation software focuses on solving in the recognition process. Feature extraction is the most basic content of modern speech recognition systems, which can effectively extract the features of English language, send precise language signals to translators, and improve the accuracy coefficient of computer translation work. The speech recognition system requires matching corresponding modules to assist teachers and students in language translation and reduce the probability of errors during the translation process.

The design of the speech system can effectively achieve educational informatization, solve problems in English teaching process, and promote students' understanding of English knowledge. After implementing speech recognition, the translator can automatically simulate training operations, thereby creating a virtual training platform. The simulation training technology uses the human-machine integration concept to achieve design, effectively combining speech recognizers and translators, effectively recognizing and judging the level of English occurrence, and can also adjust the way students' speech is targeted[4].

3 Design of a Speech Recognition System for English Translators

3.1 Hardware Design

In the design process of the line of sight English translator speech recognition system, a Samsung controller and the structure of the system's ARM processor core are used. The audio interface chip uses encoding and decoding chips, which are relatively inexpensive and can support three-wire control standards. It is the most commonly used full duplex audio chip in embedded systems. Due to the timeliness of voice signal processing and the large number of recorded and played voice signals, if the first in, first out queue is used for buffering when sending and receiving voice signals, but the data is transmitted to FIFO by the terminal, the system cost is high, and reliable recording and playing of sound cannot be guaranteed. Therefore, it is necessary to use DMA method to achieve audio recording and playing. By using this method to record and play data, the destination address, data source address, and length can be set, and the buffer can be automatically sent for filling[5]. It is not until the specified length of data is achieved that an interrupt can be applied to the system. By creating multiple buffers in memory, audio data can be effectively recorded and played.

(1) Peripheral circuit

Flash memory is widely used in various systems due to its ability to achieve electric erasure in the system, ensuring that information is not lost after power loss, and its large capacity. In this article, the internal integrated controller of the system is designed, using Nand memory with high performance, with a data storage capacity of 64MB, and using block page storage management. SDRAM chips have a relatively large capacity, low cost, and fast access speed, and are widely used in crisis management systems. SDRAM can store variables and code, referring to the memory that the system accesses after startup[6]. Because SDRAM needs to be able to refresh regularly to ensure the accuracy of stored data, microprocessors require refresh control logic. This article uses the S2C2410 microprocessor chip to implement the settings, and uses Samsung chips to create a memory system based on the actual needs of the speech recognition system.

(2) Ethernet controller

Ethernet generally uses local area network technology to quickly form a network and connect to computers through an external network. In the process of continuous development of external networks, various Ethernet control chips are also constantly emerging. This article uses CS5989 as an Ethernet controller, which can provide feasible solutions for portable products and embedded application systems. The main characteristics of this chip are low usage cost, low power consumption, good performance, and relatively powerful functions. The data transmission mode, physical layer interface, and working mode of the chip can be adjusted according to actual needs, using internal registers to achieve settings for different application environments.

(3) Serial circuit interface

It has three independent sets of serial ports for the processor unit, and this article uses two sets. The RS252 serial interface can be used when the communication distance is less than 15m. During the working process, the internal data is sent to the FIFO queue by the sending unit

using the parallel bus, and then sent through the phase shifter. In order to achieve compatibility with computer universal serial ports, a level conversion chip can also be used to convert the level into a signal that can be compatible with ordinary serial ports and communicate with peripherals[7].

3.2 Module Design

(1) Speech collection and preprocessing module

The main function of this module is to collect voice signals, and to achieve signal filtering, endpoint detection processing, and data normalization. The speech collection module is designed using WM8731 in the DE2 board. After being set using the I2C bus, it can work in the setting mode. The speech collection function structure is shown in Figure 2.

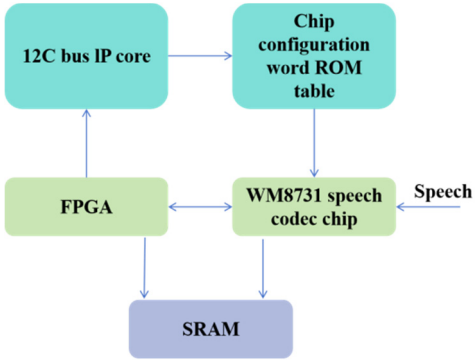

Figure. 2 Speech collection function structure

This collection unit mainly includes a PLL, I2C bus controller, and voice collection controller. The system utilizes the voice collection module to convert the sound data string collected by the voice chip into a 16 bit PCM code, which is then transmitted and saved in the memory. This achieves the setting of 4S recording time. Users input three isolated words at once, store the processed data in subsequent memory addresses, and store the detected words in the first address.

The preprocessing module mainly includes three parts, namely speech normalization processing, filter operation, and speech endpoint detection. If normalization processing is implemented according to this standard, it will waste two hardware multiplier resources. In order to achieve resource conservation, a simplified normalization processing model can be used, with the workflow as follows:

Firstly, achieve the maximum value MAX (DATA) of voice data for reading;

Secondly, make $m \geq$ MAX (DATA) and find the minimum value n;

Thirdly, the normalization operation DATA/MAX (DATA) can be transformed into DA-TA/m, which can be achieved in digital circuits with only a simple shift.

By modifying the normalization operation, it runs faster, consumes less resources, and sacrifices the sampling progress to consciously adjust the maximum amplitude for control during the signal sampling process. Because the high-frequency component of the 16 bit

original input voice is smaller than the low-frequency component, it is necessary to shift each voice data by 8 bits to the right. When n<8, there is no need to shift. After normalization, it is possible to increase the high frequency and make the signal spectrum smoother, usually using a first-order high-pass filter[8].

(2) MFCC feature extraction module

In order to effectively promote the speed of voice data processing, the entire feature extraction module is written using VERILOG as the IP core. During each MFCC operation process, data is input, and this module is a bus slave device that uses DMA to transport data. In the process of feature extraction, it is required to use fixed-point FFT design for implementation, with a focus on address generation logic and unit design. The unit design mainly includes the main computing unit of the FFT processor, which mainly uses 8 hardware multipliers to achieve data real part multiplication and data imaginary part multiplication. The butterfly structure is designed using a pipeline, with input data and output implemented in a six hour cycle. Due to the continuous input of this operation data, it is used as a clock cycle for butterfly junction calculation. The focus of the FFT module design process is on address generation logic, which can affect the operational structure and module performance. So in the implementation design process, it is necessary to use a standard address jump method and ensure that it can meet the actual operational requirements.

(3) DTW identification module

The implementation process of the module: Firstly, read the voice template library from the SD card, and use special processing to convert the voice parameter template into a binary file. Secondly, the feature parameter extraction module component can effectively obtain a real-time collection and testing module through processing, which belongs to the voice feature vector; Thirdly, by comparing the similarity between T and R, even if the accumulated distance is calculated, if the distance is small, the similarity will increase. This distortion distance can be achieved using dynamic programming[9].

4 Experimental testing

4.1 Feasibility Analysis of Translators

When conducting feasibility analysis of multi-level speech automatic recognition and translation software based on the Internet, the experimental environment is crucial. The following are some key parameters and components of the experimental environment: Speech signal sample length: In this experiment, the sample length of the speech signal was 1200. This means that each speech signal sample contains 1200 sampling points. This sample length is usually sufficient to capture typical speech segments for subsequent analysis and processing. Sampling frequency: The sampling frequency refers to the number of times a sound signal is sampled within one second. In this experiment, the sampling frequency is 24kHz, which means 24000 samples of the sound signal are taken per second. This high sampling frequency can effectively capture the details and spectral characteristics of sound signals. The carrier frequency of the translator: The carrier frequency of the translator is 30MHz. This is the carrier frequency used in the transmission and reception of voice signals. This frequency is used to modulate and demodulate the transmitted sound signal. Tuning fork as conductor: In

the experiment, a tuning fork with a frequency of 239kHz was used as the conductor. This tuning fork is used to generate sound signals as the source of input for voice signal acquisition. A tuning fork is a common experimental tool used to generate sound signals with known frequencies and amplitudes. The test model parameters are shown in Table 1.

Table 1 Test model parameters

Translation object sample	Phonetic intensity /dB	Noise intensity /dB	Filter coefficient
1	53.0910	5.7699	0.1455
2	52.2146	5.4755	0.1252
3	55.0662	5.6788	0.1488
4	54.2067	5.6872	0.1244
5	53.2435	5.7707	0.1448
6	55.3662	5.6419	0.1237
7	53.0957	5.4656	0.1402
8	54.8530	5.8547	0.1244
9	52.3906	5.6220	0.1320
10	53.8586	5.8180	0.1423

4.2 Comparison and Analysis of Speech Recognition Accuracy

To further verify the speech recognition accuracy of the designed English translator, a comparative experiment was conducted using traditional method 1 and traditional method 2 as control methods.

Use the Eclipse integrated development environment to debug the software structure of the translation system. Set up a function on the main program page to detect the pronunciation engine of the translation system, and call the testing function of the translation system. Use the Mel spectrogram generation module to frame, add windows, and preprocess the audio for speech recognition before generating Mel spectrograms. Use the obtained Mel spectrogram to reconstruct the speech waveform received by the system, save it as an audio file in wav format, and form the speech translation process. The accuracy index of speech recognition is expressed as:

$$\mathrm{STE} = (1 - \frac{\mathrm{SE}}{\mathrm{N}}) \times 100\% \tag{1}$$

In the formula, SE is the number of sentences in the identified sequence that were incorrectly identified, and N is the total number of sentences in the standard sequence. If there is a word recognition error in a defined sentence, it is considered a recognition error. The accuracy results of speech recognition obtained by three speech translation systems are shown in Table 2.

Table 2 Accuracy of speech recognition in different systems/%

Test set	System in this paper	Traditional method 1	Traditional method 2
Test set 1	89.6	79.3	61.4
Test set 2	92.7	76.2	62.9
Test set 3	90.1	77.5	65.8

Test set 4	93.4	78.2	68.3
Test set 5	88.5	73.1	60.7
Test set 6	92.3	74.7	66.1
Test set 7	89.4	76.0	67.4
Test set 8	93.9	78.8	70.1
average value	91.24	76.73	65.34

From the above table, it can be seen that the average speech recognition accuracy of the designed translator is about 91.24%, the average speech recognition accuracy of the traditional method 1 translation system is about 76.73%, and the average speech recognition accuracy of the traditional method 2 translation system is about 65.34%. In contrast, the designed translator has a higher speech recognition accuracy, indicating its strong speech recognition ability, which in turn leads to better translation results[10].

5 Conclusion

At present, in the development process of the modern information age, English teaching has gradually realized informatization. In this context, translation systems are particularly important. The speech recognition system designed in this article is the main device of the translator tool, which can effectively meet the software execution requirements and achieve the creation of translation processing processes. In the system, model training, feature extraction, etc. are the main technologies that can assist students' learning and teacher teaching, and cultivate students' English translation skills. The experimental results indicate that the designed translator has good stability and reliability, indicating its feasibility. Compared to other speech translation systems, the designed translator has a higher speech recognition accuracy and good practical application value.

References

[1] Li, R. , Wang, G. , Dai, W. , Zan, X. , & Zhang, T. . (2021). Design of distribution equipment monitoring system based on internet of things and multi-agent. Journal of Physics: Conference Series, 2093(1), 012040-.

[2] Zhang, H. , Zhu, L. , Xu, S. , Cao, J. , & Kong, W. . (2021). Two brains, one target: design of a multi-level information fusion model based on dual-subject rsvp. Journal of Neuroscience Methods, 363(5), 109346-.

[3] Mahto, K. K. , Pal, P. K. , Das, P. , Mittal, S. , & Mahato, B. . (2022). A new design of multilevel inverter based on t-type symmetrical and asymmetrical dc sources. Iranian Journal of Science and Technology, Transactions of Electrical Engineering, 47(2), 649-657.

[4] Ban, H. , & Ning, J. . (2021). Design of english automatic translation system based on machine intelligent translation and secure internet of things. Mobile Information Systems, 2021(3), 1-8.

[5] Lei, L. , & Wang, H. . (2022). Design and analysis of english intelligent translation system based on internet of things and big data model. Computational intelligence and neuroscience, 2022(3), 6788813.

[6] Shimizu, S. , Chu, C. , Li, S. , & Kurohashi, S. . (2022). Cross-lingual transfer learning for end-to-end speech translation. Journal of Natural Language Processing, 29(2), 628-637.

[7] Dhanjal, A. S. , & Singh, W. . (2022). An automatic machine translation system for multi-lingual speech to indian sign language. Multimedia tools and applications,9(3), 81.
[8] ZuchengHUANG, MengyuanSHEN, ZhichengHOU, Andrewtokuyasu, T. , & HailinMENG. (2021). Design and implementation of metabolic pathway based on multi-path breadth-first searching algorithm. Journal of Integration Technology, 10(05), 72-79.
[9] Emroozi, V. B. , Roozkhosh, P. , Modares, A. , & Kazemi, M. . (2023). A new supply chain design to solve supplier selection based on internet of things and delivery reliability. Journal of Industrial and Management Optimization, 19(11), 7993-7999.
[10] Kai-Tao, H. E. , Zhi-Zhong, L. I. , & Da-Ming, W. . (2022). Overview on the design of the service and management system for field geological survey based on the remote sensing and beidou satellites. Journal of Geomechanics, 18(3), 203-212.