Research on English Translator Speech Recognition System Based on Deep Learning

Zhiyu Zhou(✉)

Shandong Management University, Jinan 250357, China
zhouzhiyu54545@yeah.net

Abstract. The application of English translators is affected by speech recognition technology. Current speech recognition systems use Hidden Markov Models for recognition, which are susceptible to interference from noise and the magnitude of the recognition object, resulting in low accuracy and efficiency of the recognition system. Aiming at the above problems, research and design an English translator speech recognition system based on deep learning. Design the system hardware support software function with the combination of FPGA and STM32F4 microprocessor as the core. After preprocessing the speech sequence collected by the English translator and extracting the features, the DNN neural network trained by the restricted Boltzmann machine is used to recognize the speech sequence features to realize the speech recognition function. In the experiment, the DNN neural network has better recognition performance than the HMM model. The designed recognition system takes an average of 23.5 ms to recognize and has a higher recognition efficiency.

Keywords: Deep learning · English translator · Speech recognition · System design · FPGA · DNN neural network

1 Introduction

English is the most widely used language in international communication. In recent years, with the continuous development of machine translation related technologies, the application scope and demand of English translators have gradually expanded. There will be recognition errors in the use of English translators, resulting in differences in translation and affecting the use effect of translators. Therefore, for English translators, the recognition accuracy of input speech is very important. As the most important and convenient way of information exchange between people, voice is also an ideal bridge between people and intelligent hardware. Speech recognition is to let the machine understand the language spoken by human beings and convert it into accurate text information [1]. As an important interface of human-computer interaction, speech recognition has changed people's life in many aspects. Speech recognition brings a lot of convenience to production, life, work and study. The traditional speech recognition based on statistical model method uses hidden Markov model as an acoustic model component, and uses

W. Fu and G. Sun (Eds.): eLEOT 2022, LNICST 454, pp. 537–548, 2022.
https://doi.org/10.1007/978-3-031-21164-5_41

Gaussian mixture model to describe the probability of speech acoustic features. However, hidden Markov model belongs to a typical shallow learning structure. It only contains a network structure that can convert the original input signal into a specific problem space. It has a series of disadvantages, such as complex data annotation, high requirements for application scenarios and poor anti noise ability. The multi-core learning and projection algorithm in reference [2] can effectively classify multi-band noise according to different bandwidth, strengthen the speech feature level, and complete multi-band anti noise speech recognition together with CHMM model. Literature [3] produces phoneme features with different durations, and multi-core convolution fusion network is used to standardize phoneme features with different lengths and reduce the error rate of recognition words. However, the above two systems have a long time for English translator speech recognition, and the recognition accuracy is low.

With the development of deep learning, speech recognition system has achieved better and better results. The speech recognition system based on deep learning can better learn the abstract features of data, so the design difficulty of front-end feature extractor is greatly reduced, and there is no need to manually design complex feature extractor to obtain speech features [4]. The acoustic model based on deep learning needs a deep network structure. However, each voice contains hundreds of frames. With the increase of network level, more and more parameters need to be trained, and the requirements for hardware are higher. English translators have higher requirements for speech data, resulting in long delay and low recognition accuracy of the current speech recognition system. Therefore, aiming at the defects of the current speech recognition system, this paper optimizes the hardware and software, and studies and designs the English translator speech recognition system based on deep learning.

2 Research on the Hardware Part of the Speech Recognition System of English Translator Based on Deep Learning

The hardware of the English translator speech recognition system based on deep learning designed in this paper is mainly composed of speech recognition module, control module, communication module and peripheral circuit. Among them, the control module takes the STM32F4 microprocessor as the core, and consists of a power supply circuit, a crystal oscillator circuit, a reset circuit, a JTAG circuit, and a bootstrap mode selection circuit. The speech recognition module takes FPGA processing chip XC3S500E as the core, and realizes relevant data including data preprocessing, speech feature extraction and template management on the audio information collected by the microphone of the English translator. The communication module uses the integrated USB and UART communication serial ports of the control chip to upload and download data. RS-232 standard interface (also known as EIA RS-232) is one of the commonly used serial communication interface standards. Flash is an excellent web animation design software launched by Macromedia in June, 1999. It is an interactive animation design tool, which can integrate music, sound effects, animation and innovative interfaces to produce high-quality web page dynamic effects. Figure 1 below is a schematic diagram of the hardware part of the English translator speech recognition system designed in this article [5].

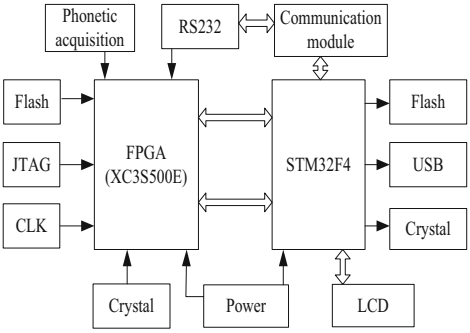

Fig. 1. Schematic diagram of the hardware framework of the speech recognition system of the English translator

2.1 Speech Recognition Module Design

In the speech recognition module, the A/D sampling of speech signal is realized through the serial port software of UDA1341TS speech chip. The frequency is 8 kHz, and the speech recognition results are displayed on LCD. The clock signal of the speech recognition module is generated by an external 32768 Hz crystal oscillator.

Considering that the software part of the speech recognition system uses deep learning algorithms for recognition processing, the data quantization of the FPGA chip uses double-precision floating-point representation. The manifestation is shown in Fig. 2 below [6].

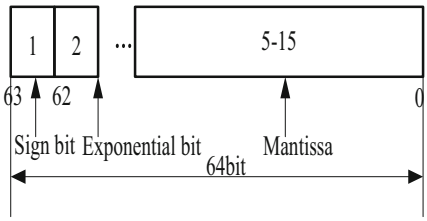

Fig. 2. Schematic diagram of double-precision floating-point representation

The 16-bit double-precision floating-point representation method is adopted. The highest bit of the voice signal is 15th for the sign bit, 14th to 12th are integer bits, and 11th to 0th bits are decimal places. In the FPGA implementation process, the specific data corresponding to the decimal operation needs to be enlarged by 212 times, and the result after the operation is correspondingly reduced by 212 times, that is, the data quantization process in the FPGA is completed. The FPGA chip is debugged through the ACT8990 interface chip. JTAG is mainly used for chip internal testing and system simulation debugging. In this design, for the convenience of online debugging, 20-pin JTAG is selected as the debugging interface. The pin function comparison is shown in Table 1.

Table 1. JTAG pins and corresponding function setting parameters

ACT8990 pin number	Name	Pin setting function
PC10	TMS	Select JTAG test mode
K11	TCKI	Test clock
L9	TDI1	Test data input
H10	TD0	Test data output
F10	RESET	Test reset

After the user releases JTAG after debugging, the GPIO controller will gain control again. After the GPIO control register is reset, the software can use these I/O ports as ordinary I/O.

The system mainly uses Flash memory to store the written related voice program code, some user data that needs to be saved after the system is powered off, constant tables and the results of voice training. This article chooses the 64M Nand Flash-K9F1208U0M chip, which can be automatically erased during programming. The voice recognition module communicates data through the RS232 serial port. When sending data, the parallel bus sends the internal data of the system to the sending unit, then enters the FIFO queue, and then passes through the sending phase shifter and then sends it through the TXDn pin.

2.2 Control Module Design

The STM32F4 microprocessor mainly has three modes of power-on, low voltage and watchdog reset. Among them, the low-voltage reset method means that when the power supply voltage is lower than a certain value, the logic judgment of the entire operating system will be disordered. In order to avoid this phenomenon, a reset signal is generated within 4 clock cycles to make each chip the value of the register returns to the initial state. The watchdog reset is to clear the watchdog at regular intervals. If it is not cleared for more than the specified time, a reset signal will be generated. The operating voltage of the STM32F4IGT6 chip is 2–3.6 V, and 3.3 V is selected as the chip's power supply. The design chooses the linear regulator AMS117 to realize the voltage conversion from 5 V to 3.3 V. The output terminal of the voltage regulator chip is connected to 0.1 μF and 10 μF capacitors, and the input terminal is connected to the same 0.1 μF and 10 μF capacitors [7]. Prevent chip damage caused by voltage inversion at the moment of power failure, and rectify the input voltage.

For the FLASH storage of the control module, the Flash memory chip selected by this system is SPR4096, which has 512 K Flash, 256 sectors of 2 K bytes, and the maximum operating frequency is 5 MHz. The STM32F4 microprocessor selects the memory mode through the combination of pin level signals. The specific memory mode selection pin level combination is defined in Table 2 below.

The microcontroller realizes the communication of the host through its asynchronous serial port UART, and displays the speech recognition results of English translation

Table 2. FLASH storage mode selection pin definition

BOOT0	BOOT1	Storage mode	Storage
0	0	SPR4096	Choose to access the main FLASH chip storage space
1	0	Chip RAM	Select the RAM storage space embedded in the microcontroller
1	1	De-expandable memory outside the system	Select system external expansion memory space

in the display information. Under the hardware framework of the speech recognition system designed above, using deep learning related algorithms, this paper analyzes the requirements of current English translators for speech recognition, designs the software part of the system, and realizes the recognition function of translated speech.

3 Research on the Software Part of the Speech Recognition System of English Translator Based on Deep Learning

3.1 English Translator Speech Signal Preprocessing

When the voice signal passes through the human glottis, it will be affected by the glottal air flow, and will be attenuated at an attenuation rate of 12 dB per octave; when it passes through the oral cavity, it will be affected by the lip radiation and will increase at an increase rate of 6 dB per octave.. The entire voice signal is attenuated by 6 dB per octave, which causes the attenuation rate to become faster and faster as the frequency continues to rise, which causes the high-frequency signal to be attenuated by a large margin. In order to reduce the energy loss caused by the attenuation of high-frequency signals, it is necessary to pre-emphasize the speech signal. Pre-emphasis can filter out low-frequency interference and improve the resolution of high-frequency components in the voice signal. The pre-emphasis operation generally passes the speech signal through a first-order high-pass filter with a characteristic of $(1-\alpha Z^{-1})$, which is called a pre-emphasis filter in most cases.

The transfer function of the pre-emphasis filter is [8]:

$$H(Z) = 1 - \alpha Z^{-1} \tag{1}$$

Among them, α represents the pre-emphasis coefficient of the speech signal, and its value range is 0.9–1.0. This article selects the pre-emphasis coefficient to be 0.98.

Speech signal is transformed with time, but the spectral characteristics of speech signal will not change in a short time. Therefore, in the process of speech signal processing, a whole segment of speech signal needs to be divided into several segments, that is, framing processing. Before speech signal feature extraction, it needs to be overlapped and segmented, that is, framing operation. When dividing frames, select 25 ms for each frame and l0 ms for frame shift, then the overlapping part is 15 ms.

In this paper, Hamming window function is used to process speech signal by windowing and framing. Hamming window function is as follows:

$$Win(k) = \begin{cases} 0.54 - 0.46\cos\left[2\pi n(K-1)^{-1}\right], 0 \le k \le K-1 \\ 0, \quad else \end{cases} \tag{2}$$

In the above formula, K is the length of the English translator's speech signal frame number. After the signal is divided into frames, the features of the speech signal are extracted to facilitate speech recognition. When extracting the characteristic parameters of the spectrogram, the discrete Fourier transform is performed on each frame of the speech signal to calculate the frequency spectrum $X(t)$, then the square is taken to calculate the energy spectrum $|X(t)|^2$, the logarithmic energy spectrum $\lg|X(t)|^2$ is calculated, and finally the energy spectrum corresponding to each frame of speech signal Rotate and splice into a matrix of eigenvectors. The discrete Fourier transform formula is shown below [9].

$$X(t) = \sum_{k=0}^{K-1} x(k)\,\omega(k) \tag{3}$$

$$\omega(k) = e^{-j2\pi K^{-1}} \tag{4}$$

In the above formula, $x(k)$ is the limited-length discrete speech signal of the English translator. After the characteristic matrix of the energy spectrum of the speech signal is obtained, the Mel frequency cepstrum coefficient of the signal is extracted. If the frequency of the speech signal of the translator is f and the mel conversion frequency is M_f, the formula for obtaining the mel frequency by using the frequency conversion of the speech signal is as follows:

$$T_M(f) = 2595\,\lg\left(\frac{f+700}{700}\right) \tag{5}$$

After converting the frequency of the speech signal into Mel frequency according to the above formula, the MFCC feature of the signal is extracted. After preprocessing the speech of the English translator, the deep learning algorithm is used to recognize the speech signal.

3.2 Deep Learning to Recognize Speech Signals

This design uses deep learning neural network to realize the speech recognition function of English translator. The multi-layer structure of the DNN neural network model can express complex functions with a small number of parameters, which is convenient for training and recognition. The DNN neural network model has a total of L + 1 layers, where the 0th layer is the input layer, the 1st to L-1 layers are hidden layers, and the Lth layer is the output layer. The adjacent layers are connected by a feedforward weight matrix. If the input vector of the DNN neural network is $I^{(l)}$ and the output vector is $O^{(l)}$, when the characteristic sequence of the speech signal collected by the microphone of

the English translator is t, the relationship between the output and input of the network is as follows:

$$I^{(l)} = \beta^{(l)} O^{(l)} + b^{(l)}, O^{(0)} = I^{(0)} = t \tag{6}$$

The activation function of the neural network uses the softmax function. The speech feature sequence t is first sent to the input layer of the 0th layer, and then propagated to each node in the hidden layer according to the arrow connection of each node, and finally to the output of the l layer Layer, and finally get the network output. For the DNN neural network, a restricted Boltzmann machine is used to train the network parameters. The restricted Boltzmann machine is composed of a hidden layer and an observation layer. Among them, the observation layer is represented by g, and the hidden layer is represented by y. The internal nodes of the hidden layer and the observation layer are independent of each other and obey the 0–1 distribution, and there is no correlation, but they are connected to each other through the weight matrix to maintain the inter-layer relationship.

For a restricted Boltzmann machine RBM, assuming the model parameter is $E = (\omega, \mu, \tau)$, its energy function is defined as [10]:

$$E(\mu, \tau) = -\sum_{i=1}^{G}\sum_{j=1}^{Y} \omega_{ij} g_i y_j - \sum_{i=1}^{G} p_i g_i - \sum_{j=1}^{Y} p'_j y_j \tag{7}$$

In the above formula, G is the number of nodes in the observation layer; Y is the number of nodes in the hidden layer; ω is the weight matrix connecting the two layers [11]; p and p' are the bias vectors of the observation layer and the hidden layer, respectively. Since the RBM structure is symmetrical, when the state of the hidden layer unit is determined, the activation states between the visible layer units are also conditionally independent [12].

Therefore, the joint probability distribution of variables μ and τ is:

$$F(\mu, \tau) = \frac{\exp(-E(\mu, \tau))}{\sum_{\mu}\sum_{\tau} \exp(-E(\mu, \tau))} \tag{8}$$

Suppose μ is the speech feature sequence of the English translator, and the edge probability density distribution of the speech feature μ is:

$$F(\mu) = \frac{\exp(-E(\mu, \tau))}{\sum_{\tau} \exp(-E(\mu, \tau))} \tag{9}$$

Since the speech feature sequence is continuous, all nodes obey Gaussian distribution by default. The task of learning RBM is to obtain the optimal solution of parameter $E = (\omega, \mu, \tau)$, so as to fit the training data [13]. In the process of fitting, it is necessary to use Gibbs sampling method to simulate and solve all voice features. In the k-step Gibbs sampling, the larger the k value, the more accurate the fitting voice feature, and the sampling process will take longer to complete. In order to obtain the optimal solution of parameter $E = (\omega, \mu, \tau)$, the maximum value can be solved by the contrast divergence

algorithm. That is, with each training data as the initial state, only k steps of Gibbs sampling are required to obtain a sufficiently good approximation. During the training process of the model speech feature sequence, due to the large number of layers of the DNN neural network model, the number of nodes in each layer is also large, and the model processes a large number of feature sequences, it is inevitable that over-fitting problems will occur. Therefore, the Dropout strategy is introduced to suppress the over-fitting problem in the model, so as to ensure the accuracy of the final recognition result of the model [14]. In the parameter tuning process, by setting all nodes in each hidden layer of the model to 0 with a certain probability during the training process, the optimized model is averaged to avoid output over-fitting. Perform parameter training on the DNN neural network according to the above process to determine the translator's speech recognition parameters. Use the trained DNN neural network to recognize the input speech signal sequence. The above is the design process of using the deep learning algorithm to realize the speech recognition software function part of the English translator. The software part is transplanted to the hardware, and the design and research of the speech recognition system of the English translator based on deep learning is completed.

To sum up, the overall process of the English translator speech recognition system based on deep learning designed in this paper is shown in Fig. 3.

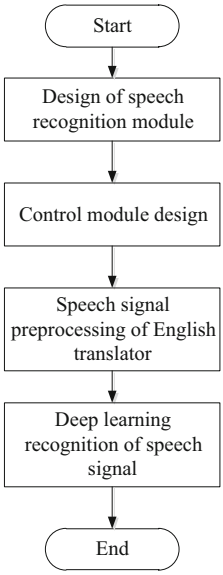

Fig. 3. Overall process of the English translator speech recognition system based on deep learning designed in this paper

4 Experimental Research

Before applying the above designed English translator speech recognition system based on deep learning to practice, it is necessary to test the system in all aspects to ensure the normal functions of the system, and study the performance of the system.

4.1 Experiment Content

The experiment is divided into two parts. One part tests the comprehensive operation performance of the system, and the other part tests the performance of the speech recognition model. The speech recognition system based on Hidden Markov model is compared with the recognition system designed in this paper.

In the performance experiment of the speech recognition model, the Switchboard and RT03S speech library are selected as the training test set. Among them, the Switchboard voice library contains 4870 conversations from 520 speakers, and about 309 h of voice data. Among them, about 30 min of voice data is selected as the test set; and the RT03S voice library contains Switchboard and Fisher. Set, respectively select the speech data of about 30 min in these two sub-data sets as the test set. Use hidden Markov model and DNN neural network model for parameter training, and recognize with the speech sequence in the test set. The performance of the model can be intuitively compared by comparing the loss of the model and the accuracy of the recognition in the recognition process.

In the system performance experiment, two speech recognition systems are applied to the English translator. In order to avoid interference, recording equipment is used as the input of English translator. By comparing the recognition time and accuracy of the system, combined with the experimental analysis of speech recognition model, the final conclusion is drawn.

4.2 Experimental Results

Table 3 below shows the comparison of system recognition time-consuming and accuracy of the two speech recognition systems when recognizing the speech signal input into the English translator in the system comparison experiment.

It can be seen from the data in Table 3 that the recognition time of the system in this paper is lower than that of the comparison system, and the recognition accuracy of the system is higher than that of the comparison system. Further processing the data in the table, the average recognition time of the system in this paper is 23.5 ms, and the average recognition time of the comparison system is 69.84 ms. The recognition accuracy of this system is up to 97.8%, while that of the comparison system is up to 94.7%. The recognition efficiency of this system is higher and the recognition effect is the best.

Figure 4 and Fig. 5 show the comparison of recognition loss and accuracy when the model identifies the test set.

Analyzing Figs. 4 and 5, it can be seen that when the HMM model recognizes the speech sequence in the training set, the recognition loss is higher than that of the DNN neural network model. At the same time, the overall recognition accuracy of DNN

Table 3. Comparison of system identification time and accuracy

Number of recognition speech sequences	Text system		Comparison system	
	Recognition time/ms	Recognition accuracy rate/%	Recognition time/ms	Recognition accuracy rate/%
25	22.2	96.8	38.3	94.7
50	22.5	96.9	39.7	94.5
100	22.8	96.1	45.3	94.5
150	23.2	96.5	56.1	94.3
200	23.5	97.8	67.5	91.6
250	23.6	97.2	74.6	91.7
300	23.9	97.8	80.2	89.0
350	24.3	96.7	88.7	89.2
400	24.4	96.5	96.4	89.4
500	24.6	96.2	111.6	89.1

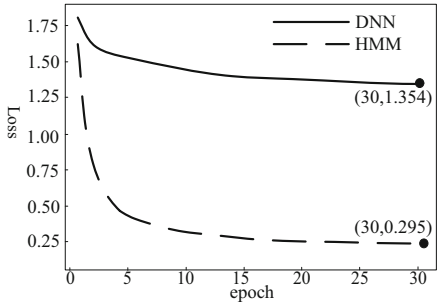

Fig. 4. Comparison of speech recognition model loss

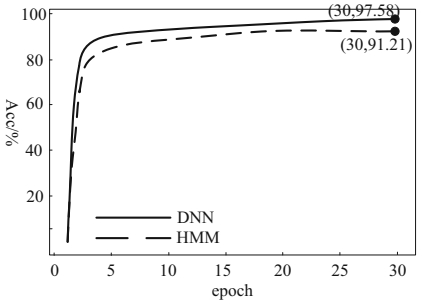

Fig. 5. Comparison of accuracy of speech recognition models

neural network model is higher than that of HMM model. It can be seen from the model experiment that the performance of the DNN model used in this paper is better in speech recognition.

5 Conclusion

Voice is the most commonly used and convenient information carrier for human communication and information sharing. It is also the most natural and indispensable information medium in human-computer interaction. English translator is an important achievement of machine learning, which effectively improves the convenience of English communication. However, the translation accuracy of English translator is also limited by speech recognition technology. In recent years, speech recognition technology triggered by deep learning has promoted the development of speech recognition related fields. Aiming at the defects of traditional speech recognition system, this paper studies and designs an English translator speech recognition system based on deep learning. The DNN neural network trained by restricted Boltzmann machine is used to recognize the features of speech sequence and realize the function of speech recognition. Through the test of the system, it is determined that the designed system can meet the requirements of the current English translator for the speech recognition system, and the recognition accuracy, speed and other performance of the system are improved to a certain extent compared with the traditional speech recognition system. However, the algorithm of this system in the process of English translator speech recognition is complex, which leads to the time of English translator speech recognition not reaching the expectation. Therefore, in the next research, the algorithm is improved to improve the efficiency of English translator speech recognition.

References

1. Wang, J. Xu, S.-L., Yu, Z.-T., et al.: Chinese-vietnamese speech translation with deep pre-encode convolutional neural network. J. Chin. Comput. Syst. **42**(04), 736–739 (2021)
2. Gu, H.-H.: Multi-band anti-noise speech recognition method simulation based on multi-core learning. Comput. Simul. **36**(10), 364–367+395 (2019)
3. Xiaofeng, L., Wenai, S., Xiaodong, C., et al.: BLSTM-CTC speech recognition based on multi-core convolutional fusion network. Comput. Appl. Softw. **38**(11), 167–173 (2021)
4. Long, Y., Li, Y., Zhang, Q., et al.: Acoustic data augmentation for Mandarin-English code-switching speech recognition. Appl. Acoust. **161**(11), 107175 (2020)
5. Hu, L., Huang, H., Liang, C., et al.: Research of end-to-end speech recognition based on double-path CNN. Transducer Microsyst. Technol. **40**(11), 69–72+83 (2021)
6. Dong, J., Li, S.: English speech recognition and multidimensional pronunciation evaluation. Educ. Res. Front. Chin. Engl. **010**(003), 184–188 (2020)
7. Xlao, X., Xu, C.: Speech feature fusion algorithm based on acoustic state likelihood and supervised state modelling. J. Tsinghua Univ. (Sci. Technol.) **59**(06), 476–481 (2019)
8. Lu, X., Shah, M.A.: Implementation of embedded unspecific continuous English speech recognition based on HMM. Recent Adv. Electr. Electron. Eng. **6**, 649–659 (2021)
9. Bai, L., Wang, L.-M.: Convolutional neural network for speech recognition. J. Northeast Norm. Univ. (Nat. Sci. Ed.) **52**(02), 52–57 (2020)

10. Li, P., Yang, Y., Gao, X., et al.: A study of Chinese speech recognition based on bidirectional recurrent neural network. Appl. Acoust. **39**(03), 464–471 (2020)
11. Wang, Q., Yan, L.: English translation method based on recurrent neural network. Autom. Technol. Appl. **39**(11), 5 (2020)
12. Sun, J., Yan, B.: Application of new media technology in English translation and translator's subjective cognition. Light Alloy Process. Technol. **48**(11), 1 (2020)
13. Li, H.: Characteristics and skills of professional English translation. Thermosetting Resin **36**(3), 2 (2021)
14. Zhang, H., Huang, H., Li, W., et al.: A review of speech emotion recognition. Comput. Simul. **38**(8), 11 (2021)