INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING
DOI: 10.46300/9106.2022.16.117 Volume 16, 2022

Application of Speech Recognition Technology in Chinese English Simultaneous Interpretation of Law

Xiao Yang*
College of Foreign Languages, Xijing University, Xi'an 710123, China

Received: August 26, 2021. Revised: February 14, 2022. Accepted: March 12, 2022.
Published: March 30, 2022.

Abstract—Speech recognition is an important research field in natural language processing. In Chinese and English, which have rich data resources, the performance of end-to-end speech recognition model is close to that of Hidden Markov Model—Deep Neural Network (HMM-DNN) model. However, for the low resource speech recognition task of Chinese English hybrid, the end-to-end speech recognition system does not achieve good performance. In the case of limited mixed data between Chinese and English, the modeling method of end-to-end speech recognition is studied. This paper focuses on two end-to-end speech recognition models: connection timing distribution and attention based codec network. In order to improve the performance of Chinese English hybrid speech recognition, this paper studies how to improve the performance of the coder based on connection timing distribution model and attention mechanism, and tries to combine the two models to improve the performance of Chinese English hybrid speech recognition. In low resource Chinese English mixed data, the advantages of different models are used to improve the performance of end-to-end models, so as to improve the recognition accuracy of speech recognition technology in legal Chinese English simultaneous interpretation.

Keywords—speech recognition, law, Chinese English simultaneous interpretation of law.

I. Introduction

SIMULTANEOUS interpretation is a simultaneous interpreting method, which is an interpretation way that an interpreter keeps translating his speech to the audience without interrupting the speaker's speech. Simultaneous interpretation has been widely used in international conferences, business activities and other fields, as well as in small-scale occasions such as journalists' interviews and banquet speeches [1]. With the rapid development of deep neural network, the speech recognition community has begun to use the deep neural network technology to deal with speech recognition tasks. The deep neural network is directly integrated into Hidden Markov Model—Gaussian Mixture (HMM-GMM) model to form HMM-DNN. The end-to-end speech recognition system is constructed directly by referring to the end-to-end idea in the field of machine translation. The end-to-end speech recognition system is relatively simple, and it does not need complicated alignment and pronunciation dictionary construction, which shows a good application prospect. However, so far, simultaneous interpretation is still dominated by artificial simultaneous interpretation [2]. Although software technology is developing rapidly, the research on automatic simultaneous interpretation technology is still in the experimental stage. At the same time, speech recognition, machine translation and Text To Speech (TTS) technology have been mature. The famous related products include Microsoft voice recognition engine, Google online translation service, Microsoft TTS voice engine, etc. In addition, most of the Application Program Interfaces (APIs) related to speech recognition, machine translation and TTS technology have been published. This provides us with great convenience in software development by using these three technologies [3]. Therefore, a service aggregation based simultaneous interpretation software can be designed to aggregate speech recognition, machine translation and TTS technology, which makes simultaneous interpretation possible. These interpreting techniques play a multiple role in enabling, assisting and empowering [4]. They provide interpreters with different forms and different degrees of help in the pre translation, during and after translation, such as obtaining professional knowledge quickly, grasping semantic abstract information, extracting professional terminology knowledge, clarifying the logic relationship of the original text, and managing the language assets of interpretation. Based on the prediction of accuracy and translation effect, many scholars have also tested. This paper analyzes the bilingual translation from the aspects of vocabulary, sentence pattern and speech recognition [5]. By comparing

the working models of the simultaneous interpreter and Machine Translation, and investigating the technical bottlenecks of Machine Translation, this paper discusses the possibility of the simultaneous interpreting system and proposes solutions from the perspective of computer-assisted interpretation.

II. Speech Recognition Technology Of Chinese English Simultaneous Interpretation Of Law

A. Speech Feature Recognition in Legal Simultaneous Interpretation

With the continuous improvement of neural network translation technology based on deep learning, the construction of automatic machine simultaneous interpretation system has the basic practical conditions. However, so far, the machine simultaneous interpretation system is still in the paper stage, and few commercial products based on machine simultaneous interpretation technology have come out [6]. Chinese English simultaneous interpretation of legal speech recognition links a variety of existing enterprise level services together to build a new enterprise level solution. Its essence is to make comprehensive use of existing services to provide the most appropriate aggregation service solution for specific business needs. The basic pattern of service aggregation is shown in Figure 1.

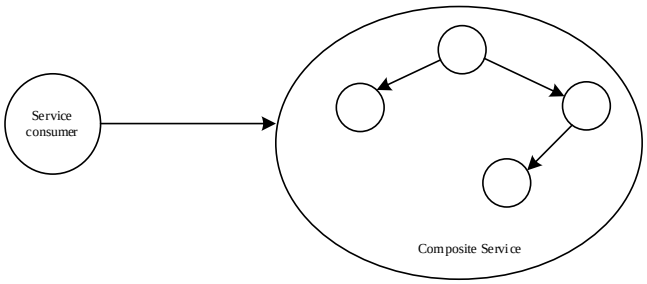

Figure 1. The aggregation mode of speech information in legal recognition

When aggregating services, there are two different execution models: orchestration and choreography. In the orchestration model, the interaction of all component services must be described, just as in a traditional workflow system [7]. The description will be executed by an orchestration engine that controls the entire aggregation. In the choreography model, instead, each component service only knows its own interaction and behavior, and no entity can control the whole service aggregation at the global level. The interface of speech recognition, machine tran slation and TTS sub aggregation services is designed to facilitate the expansion of functions and aggregation of more services [8]. Voice recognition service adopts Microsoft voice recognition engine and SAPI provided by Microsoft. The user can set up the language that is recognized. The machine translation service provides a set of interfaces, and encapsulates Google online translation service, Microsoft online translation service and Yahoo online translation service, which can be selected by users [9]. TTS service adopts Microsoft TTS voice engine and SAPI provided by Microsoft. Users can set the type of reading language. The structure of the platform is shown in Figure 2:

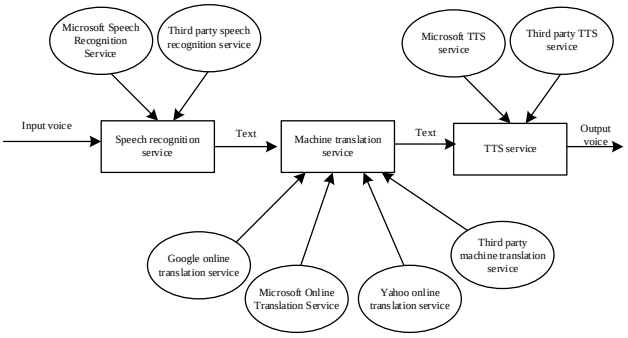

Figure 2. Structure of simultaneous interpretation platform

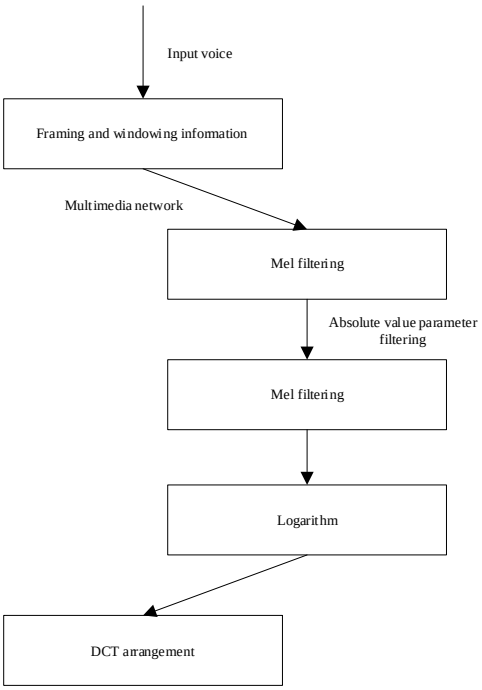

Figure 3. Schematic diagram of pitch feature extraction of Chinese and English legal speech

In speech recognition service, users can select the voice recognition sub service they want to use (the platform only provides Microsoft voice recognition service at present), then set the recognized language through the interface, input voice, and convert the voice into text and output it to the machine translation service module [10]. In MT service, users can call the machine translation sub service they want to use, translate the input text from one language to another, and then output it to TTS service module in the form of text. In TTS service, users can select the appropriate TTS sub service (the platform only provides Microsoft TTS service at present), then set the type of reading language through the interface, convert the input text into speech and output it in the form of voice. The above three services with specific functions are effectively aggregated according to the process model shown in Figure 2, and then a service aggregation platform can realize simultaneous interpretation function is obtained [11]. According to the characteristics of the constant transmission of voice fuzzy

tone data in the multimedia network array, it can be seen that when the input voice contains the information of frame adding window, the multimedia network will filter the absolute value parameters of the speech signal by means of Mel filtering, and then take logarithm processing according to the processing results, and finally make the feature information of the fuzzy voice tone present DCT arrangement. Figure 3 reflects the complete principle of extracting the tone features of fuzzy speech.

Let f be the total amount of framing and windowing information in the input speech, k be the uniform transmission rate of the multimedia network array, and s be the Mel filter coefficient:

$$Ael(f)=\lambda\ln(1+\frac{\sqrt{a^2+d^2}}{ks}) \tag{1}$$

In the formula, $Ael(f)$ represents the fixed form of absolute value parameter of fuzzy speech tone, λ represents the authority coefficient of the filter process, a, d are represent two different eigenvalue parameter points. In order to clearly express the characteristic information of fuzzy speech tone, g is used to represent the basic DCT arrangement rule, χ stands for the extraction parameter, and the simultaneous formula can express the result of fuzzy speech tone feature extraction as follows:

$$H(f)=\frac{\chi[Ael(f)-1]-k}{gl^2\Delta\delta} \tag{2}$$

In the above formula, k represents the minimum value of the absolute value parameter of the fuzzy voice tone allowed in the extraction process, and l represents the difference of representation features under the DCT arrangement rule, $\Delta\delta$ represents the change of the absolute value parameter during the extraction operation.

B. Chinese and English Legal Speech Denoising Algorithm

Speech recognition service is not only bound with Microsoft voice recognition service, but also can be bound with other specific voice recognition sub services. The implementation of speech recognition service in the project is to adopt the voice recognition engine and SAPI provided by Microsoft [12]. It is necessary to install two software packages, microsoftspeech sdk5.1.msi and Microsoft speechsdk5.1language pack.msi in the computer. The specific implementation strategy is to design the Microsoft SP recognition class, which is used to implement the interface SP recognition [13]. The Microsoft SP recognition class implements the interface method by calling the create grammar, Division set state, get registers and other APIs in SAPI. In the speech recognition framework, a hidden Markov model is needed for each phoneme. Taking Kaldi as an example, the real phoneme model consists of three states, ASR is a mathematical model established by Bayesian decision theory, among which the most likely word sequence W is estimated in all possible word sequences v^*, and the formula is as follows:

$$\hat{W}=\arg\max P(W|\ X) \tag{3}$$

Therefore, the main problem of ASR is how to obtain the posterior distribution $P(W|\ X)$. At present, the ASR system based on hybrid HMM-DNNX uses Bayesian theorem to introduce HMM state sequence s to decompose $P(W|\ X)$ into the following three distributions.

The acoustic model $P(X|\ S)$ is further decomposed by using probability chain rule and conditional independence assumption as follows:

$$\begin{aligned}P(X|\ S)&=\sum_{t=1}^{T}P\left(x_t|\ x_1,\ldots,x_{t-1},S\right)\\&\approx\sum_{t=1}^{T}P\left(x_t|\ s_t\right)\propto\sum_{t=1}^{T}\frac{P\left(s_t|\ x_t\right)}{P\left(s_t\right)}\end{aligned} \tag{4}$$

The frame by frame likelihood function $P\left(x_t|\ s_t\right)$ is calculated by using the so-called pseudo likelihood trick and powerful DNN classifier:

$$P(S|W)=\underbrace{\sum_{t=1}^{T}P\left(s_t|s_1,\ldots,s_{t-1},W\right)}\sum_{t=1}^{T}P\left(s_t|s_{t-1},W\right) \tag{5}$$

This probability is obtained by the given HMM state transfer. The transition from w to HMM is achieved by phonemes, which are determined by the pronunciation dictionary. In Kaldi, the term HMM model is actually implemented by a restricted state converter [14]. Its input label is phoneme and output label is word or word. The language model $P(W)$ is decomposed by using probability chain rules and conditional independence hypothesis (n-1 order Markov hypothesis) as N-gram model, that is:

$$\begin{aligned}P(W)&=\sum_{m=1}^{M}P\left(w_m|\ w_1,\ldots,s_{m-1},W\right)\\&\approx\sum_{m=1}^{M}P\left(w_m|\ w_{m-n-1},\ldots,w_{m-1}\right)\end{aligned} \tag{6}$$

It can be seen from the above formula that the higher the probability of the language model of the word sequence, the smaller the perplexity of the language model, and the higher the probability of the language model, the higher the confidence of the prediction of the next word, and the stronger the ability of the language model, because the perplexity is an evaluation matrix that does not depend on a specific data set [15]. Therefore, as long as the output Dictionary of the language model is the same, it can be used Confusion can be used to compare the performance of different language models. The object of speech recognition system is strong coupling words or phrases, that is, the only standard output answer after the voice waveform is input. As mentioned above, speech recognition system can understand “proper nouns or industrial terms” which some interpreters do not understand or listen to easily, as compared with human interpreters [16]. If the interpreters make proper nouns and corresponding translations into the terminology database before translating, the speech recognition system detects corresponding keywords and displays corresponding translations in real time simultaneous interpreting [17]. At this point, the interpreter

reaches the computer with the computer. Through interaction, a simple computer-aided interpretation system has been formed. As is shown in Figure 4.

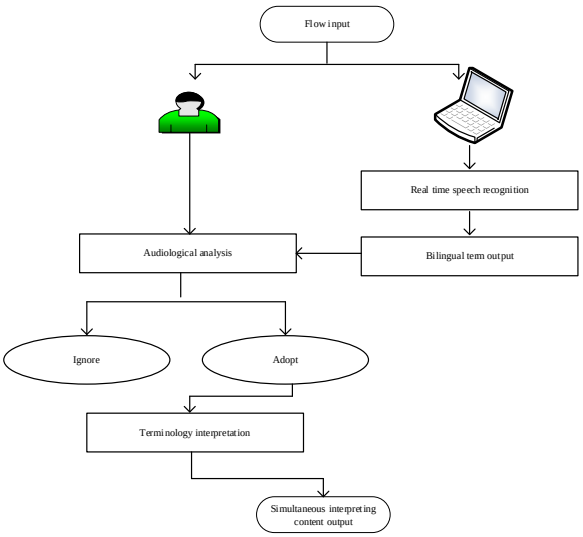

Figure 4. Speech denoising decoding process flow

The decoding network has the ability to distinguish the fuzzy tone data, and can analyze the hidden layer unit state in the data recognition operation according to the representation ability of the input speech. Multimedia network is the input medium of voice fuzzy tone data [18]. Controlling the correlation between decoding network and multimedia network is an important condition to improve the recognition accuracy [19]. Attention sub network unifies many physical parameters such as decoding authority of tone data, coding weighted output value and so on, and exchanges these data with neural network unit in the form of vector, so as to stabilize the environment of fuzzy tone data recognition. Based on this, the detailed steps are as follows. As is shown in Figure 5.

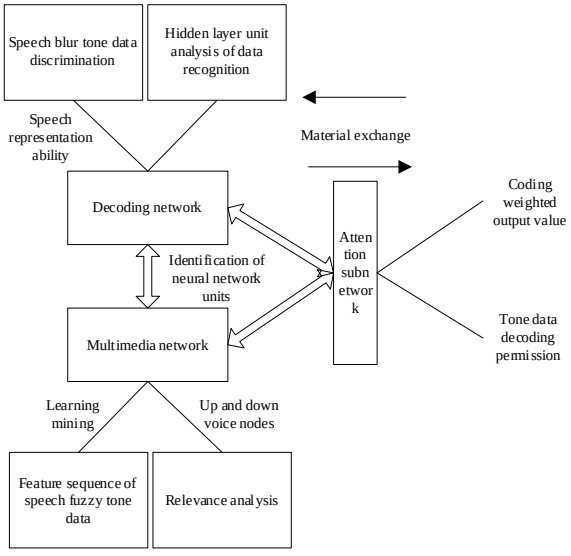

Figure 5. Optimization of voice recombination steps

The language model here is actually a restricted state receiver in Kaldi. Its input label is a word or a word, and its output label is also a word or a word, and the input label is the same as the output label. Although the recurrent neural network language model can avoid this kind of conditional independence assumption problem, it makes decoding more complex [20]-[21]. Therefore, the first decoding usually uses the n-gram language model, and RNNLM is usually used to re score the decoded lattice on the basis of the first decoding to improve the recognition results of the test set. The computer assisted interpretation system can achieve the basic computer aided interpretation function with lower cost by docking the mature speech recognition engine provided by the software service provider and the industry terminology library built by the simultaneous interpreting interpreter or the interpreter organization. For large organizations or institutions with sufficient funds and high accuracy requirements, semi-automatic computer-aided translation can be realized by building a proprietary platform combining speech recognition and machine translation system. For example, open source translation engine tensor flow based on deep neural network and open source speech recognition engine Kaldi can be used to build their own machine simultaneous interpretation by training free corpus. Generally, word error rate is used to evaluate the performance of speech recognition system, and confusion degree is used to evaluate the quality of language model. Given a word sequence W1: m, the confusion degree can be calculated as follows:

$$P_p\left(w_{1:M}\right) = P\left(w_1 w_2 w_3 \ldots w_M\right)^{-\frac{1}{M}} \tag{7}$$

In the formula, the joint probability can be decomposed into.

$$P\left(w_1 w_2 w_3 \ldots w_M\right) = P\left(w_i \mid w_{1:i-1}\right) \tag{8}$$

The conditional probability of decomposition can be obtained from the language model.

It can be seen from the above formula that the higher the probability of the language model of the word sequence, the smaller the perplexity of the language model, and the higher the probability of the language model, the higher the confidence of the prediction of the next word, and the stronger the ability of the language model, because the perplexity is an evaluation matrix that does not depend on a specific data set. Therefore, as long as the output Dictionary of the language model is the same, it can be used Confusion can be used to compare the performance of different language models.

C. Realization of English Simultaneous Interpreting Speech Recognition in Law

Suppose a speech recognition system has been trained, and now a test set is used to evaluate the performance of recognition. A common practice is to send the test set into the speech recognition system. The system has a prediction text corresponding to each audio sentence. During the preparation of the test set, it is manually proofread by people for the transcribed text of each sentence, which is proofread by people The transcribed text (called reference text) is considered to be correct, and the text predicted by speech recognition system (called hypothetical text) is compared with it. Assuming that both the reference text and the hypothetical text are based on words, the minimum

editing distance between the two texts is calculated, and the word error rate is obtained:

$$W_{ER} = \frac{S + D + I}{N} \times 100\% \quad (9)$$

In the formula S is the number of replacement errors, D is the number of deletion errors, I is the number of insertion errors, and N is the total number of words. The Chinese English hybrid speech data set used is seam, which is a microphone based spontaneous conversation bilingual speech corpus, in which most sentences contain English and Chinese. The data set consists of a training set train and two test evaluation sets. The two test evaluation sets are the words of 10 people randomly selected from the 154 people, and the test evaluation set DSE. The content of speech is mainly in English, and the test evaluation set is DMA. The main content of the speech is Chinese. The details of the whole seam data are shown in the Table 1.

Table 1. Chinese and English phonetic data sets

	Train	d_{sge}	d_{man}
Number of speakers	134	10	10
Duration	101.1	4	7.5
Proportion of Chinese	0.59	0.29	0.69

Because the code of source language and target language need to be specified when calling web service, and each language has its corresponding code, the implementation strategy of set source and set target interface method is: first convert the input language name into its corresponding code, and then assign the resulting code to the two properties representing source language and target language in the class. The implementation strategy of the translate interface method is: first set each web attribute in the class, such as the address of the web translation service and the address of the web source language detection service. After setting these properties, embed them into the URL. By creating an httpweb request object, the information is sent to the server. After that, an HTTP Web response object is obtained through the get response method. The response message is obtained by calling the member function get response stream of the object, and then the content of the response message is stored in a string by calling the read to end method. After getting the string that stores the content of the response message, the translation result can be obtained by parsing the string. The translation result is output as the return value of the translate method. The interpreter needs to prepare the meeting content in advance to expand the bilingual corpus, complete the five working links from listening comprehension to translation output in unit time (t_1-t_2) of interpretation activities, and coordinate the allocation of resources in different links in real time. If one of the links is affected in the coordination time and exceeds the interpreter's own overall ability, it is easy to lengthen the interpreter's time difference between listening and translation, and eventually lead to the delay the formation of spillover effect. Based on this, we build the working model of the simultaneous interpreting of the legal system. As is shown in Figure 6.

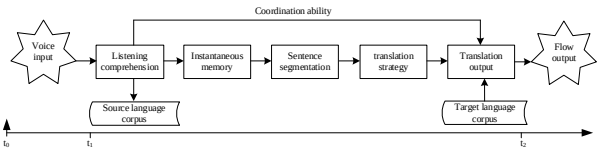

Figure 6. Working model of speech recognition in simultaneous interpreting

Although machine interpreting can imitate the output of the simultaneous interpreting interpreter through specific processes, the machine interpretation system based on probability estimation is essentially different from the simultaneous interpreting work model based on deep semantic understanding and translation strategies. The object-oriented of the existing machine translation engine is mostly the whole written text with sentence break. Before using machine translation, users with translation foundation will pre edit it to improve the quality of machine translation. However, the object-oriented of machine simultaneous interpretation system is more complex. The input text is mostly real-time speech flow in strong noise environment, and contains a lot of grammatical errors and fragments of speaker self-correction. It's a simple text. Therefore, the machine simultaneous interpretation system needs to complete the three main steps from speech recognition, machine translation to speech synthesis in unit time (t_1-t_2), as shown in Figure 7.

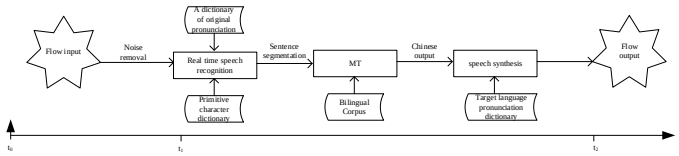

Figure 7. Process optimization of voice recognition in legal simultaneous interpreting

The simultaneous interpreting system can be divided into four parts: speech recognition, punctuation cutting, and the integration of Machine Translation and speech synthesis. Taking speech recognition as an example, the results after processing are affected by sudden noise, and the short-term average energy of some speech frames suddenly increases, which makes the recognition results inaccurate. Therefore, the processing stage flow is designed as shown in Figure 8.

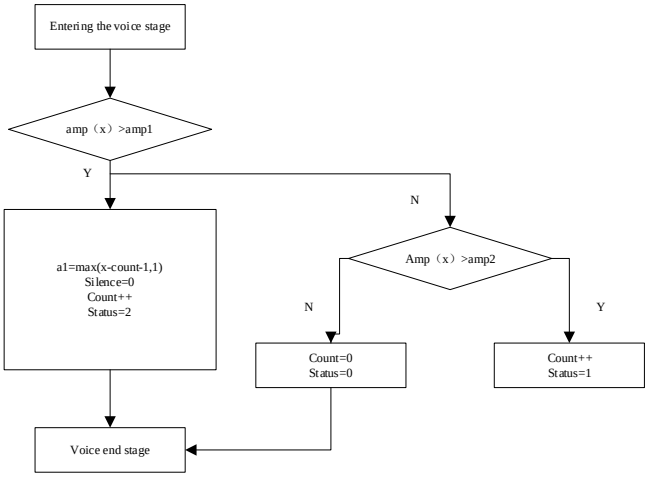

Figure 8. Process flow of Chinese and English legal speech recognition

The specific implementation steps are as follows:

When the voice signal is in the silent stage, set status = 0 to increase the voice signal frame. If the short-term energy of a frame is too high, then the frame is the starting point of the voice signal. At this time, set status = 1 to indicate that the voice signal has entered the transition period, and it is impossible to determine that this part is a voice segment.

If the short-time energy of a frame is too low, then the frame indicates that the transition period is restored to the silent stage, and the status is 0.

If the short-time energy of the frame is higher than amp1, and the frame number continues to increase, it can be determined that the signal has entered the speech stage. At this time, the status = 2, and the current speech frame number is the initial point of speech.

If the current frame is a speech segment, then status = 2, and the short-term energy of the speech frame is lower than amp2, then the segment is noise.

Continue to increase the frame number, when the duration is greater than the silence stage, it means that the voice signal endpoint is normal and can output effective voice.

According to the system software design process, 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.

Based on the above steps for simultaneous interpretation speech recognition processing can better improve the recognition accuracy and ensure the recognition effect.

III. Analysis Of Experimental Results

This experiment uses Kaldi speech recognition toolbox, which is developed by Professor Dan Povey. Kaldi is mainly divided into the bottom operation library and the upper user interface. The bottom library is mainly written by C + + language, and the upper user interface is mainly written by perl language, shell language and python language. It is an important tool in the speech recognition community at present, and the development community activity of the tool is not only widely used in academic circles, but also widely used in industry. The tool provides a large number of high-quality components, which can quickly build a speech recognition baseline system. After the development of the software, it has basically realized the simultaneous translation between English and Chinese. The development environment is based on centos7 cluster server, CPU es-2680 V4 @ 2.40GHz 28 core processor, and GPU is a single NVIDIA Tesla P100 computing card. The data of this experiment comes from the sea data set, and two test sets in the data set are used to evaluate the model. Because the Chinese English hybrid sea data set is a small data set, it needs to take the way of data augmentation to increase the amount of network training data. Therefore, the audio data of the training set is perturbed by speed and volume, and then the 80 dimensional FBANK feature and 3-dimensional pitch feature are obtained. Before being sent to the network, the audio feature is augmented by specaugment data. 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. Only 80 dimensional FBANK features and 3-dimensional pitch features are obtained from the test set data. Assuming that the distance between the source of fuzzy speech input and the central organization of context sensitive block remains unchanged, and the adjacent speech sequences can only be transmitted on demand by directional movement, the complete spectrum feature structure of Chinese English translation is shown in Figure 9.

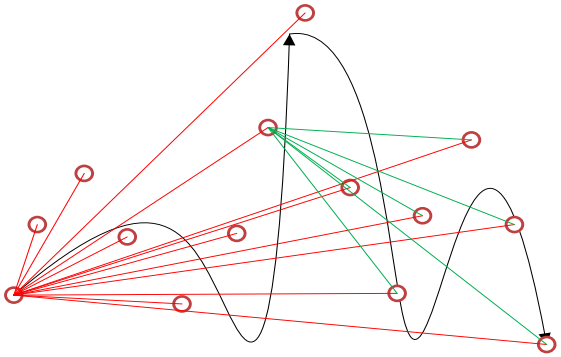

Figure 9. Structure of spectrum characteristics of Chinese English translation

The cross ordinates of simultaneous interpreting represent the different transmission directions of the context sensitive blocks, and the different node rays represent the spectrum characteristic lines of different fuzzy voice sources respectively. Under the condition that the speech cutting parameter is 0.22, 55 min is taken as the experimental time, and the change of signal segmentation rate is recorded after the application of the recognition method in the experimental group and the control group. The detailed experimental comparison results are shown in the Figure 10.

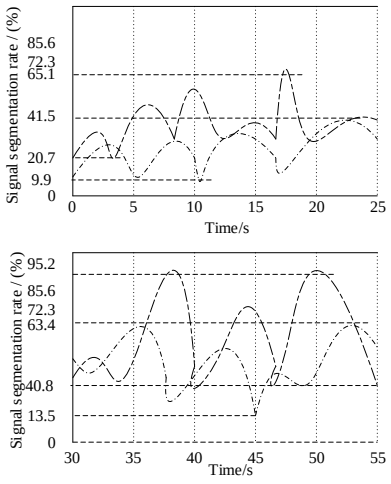

Figure 10. Speech recognition effect comparison

Further, a certain number of words, short sentences and long sentences are selected from English and Chinese for accuracy test. In order to ensure the reliability of the test results, Google translation service, Microsoft translation service and Yahoo translation service are used respectively. It is found that the results of the three tests are similar. Therefore, the final test result is the average of the three test results. The final test results are shown in Table 3.

Table 3. Accuracy test results of simultaneous interpretation platform

Test content	Accuracy (before engine training)	Accuracy (after engine training)
word	75%	85%
Short sentences	70%	77%
Long sentences	60%	63%

It can be found from the table that the simultaneous interpretation platform has the highest accuracy rate for words and the lowest accuracy rate for long sentences. Through the training of Microsoft speech recognition engine, the accuracy rate of simultaneous interpretation of words and short sentences increases greatly, while that of long sentences increases slightly. The accuracy of simultaneous interpretation of words and short sentences is mainly affected by Microsoft speech recognition service. Speech recognition engine training can improve the accuracy of interpretation.

IV. Conclusions

Based on the idea of service aggregation, the existing speech recognition and machine translation technologies are effectively combined to realize the simultaneous interpretation platform. As a product of the idea of service aggregation, the software makes full use of the idea of leading the world's technology trend in structure and technology, and organically integrates various existing open source services, so as to realize the valuable function of simultaneous interpretation. Of course, the simultaneous interpretation platform still needs to be improved in some aspects. There are three key tasks in the follow-up improvement work: automatic detection of breakpoints; encapsulation and aggregation of more services; automatic service discovery. These three improvements play an important role in improving the service quality of the software. To sum up, with the improvement of speech recognition and machine translation, the quality of interpretation will be greatly improved, which can greatly improve its practical value.

References

[1] Oh Y R, Park K, Jeon H B, et al. Automatic proficiency assessment of Korean speech read aloud by non-natives using bidirectional LSTM-based speech recognition. ETRI Journal, 2020, 42(10):59-64.
[2] Hovsepyan S, Olasagasti I, Giraud A L. Combining predictive coding and neural oscillations enables online syllable recognition in natural speech. Nature Communications, 2020, 11(1):78-84.
[3] Cabral F S, Fukai H, Tamura S. Feature extraction methods proposed for speech recognition are effective on road condition monitoring using smartphone inertial sensors. Sensors, 2019, 19(16):3481-3488.
[4] Kumar A, Aggarwal R K. Discriminatively trained continuous Hindi speech recognition using integrated acoustic features and recurrent neural network language modeling. Journal of Intelligent Systems, 2020, 30(1):165-179.
[5] Liu L, Feng G, Beautemps D, et al. Re-synchronization using the hand preceding model for Multi-modal fusion in automatic continuous cued speech recognition. IEEE Transactions on Multimedia, 2020, 12(99):1-10.
[6] Newgord C, Tandon S, Heidari Z. Simultaneous assessment of wettability and water saturation using 2D NMR measurements. Fuel, 2020, 270(11):117-131.
[7] Goerlandt F. Maritime autonomous surface ships from a risk governance perspective: Interpretation and implications. Safety Science, 2020, 128(6):104758.
[8] Mahalingam S, Bhalla N M, Mezrich J L. Curbside consults: Practices, pitfalls and legal issues. Clinical Imaging, 2019, 57(5):83-86.
[9] Shi Y Y, Bai J, Xue P Y, et al. Fusion feature extraction based on auditory and energy for noise-robust speech recognition. IEEE Access, 2019, 7(10):81911-81922.
[10] Viswanathan N, Kokkinakis K. Listening benefits in speech-in-speech recognition are altered under reverberant conditions. The Journal of the Acoustical Society of America, 2019, 145(5):348-353.
[11] Yazdani R, Arnau J M, Gonzalez A. A low-power, high-performance speech recognition accelerator. IEEE Transactions on Computers, 2019, 68(12):1817-1831.
[12] Kim G, Lee H, Kim B K, et al. Unpaired speech enhancement by acoustic and adversarial supervision for speech recognition. IEEE Signal Processing Letters, 2019, 26(1):159-163.
[13] Montenegro C, Santana V, Lozano J A. Analysis of the sensitivity of the End-Of-Turn detection task to errors generated by the automatic speech recognition process. Engineering Applications of Artificial Intelligence, 2021, 100(1):104-109.
[14] Sun R H, Chol R J. Subspace Gaussian mixture based language modeling for large vocabulary continuous

speech recognition. Speech Communication, 2020, 117(10):21-27.

[15] Martinez A C, Gerlach L, Payá-Vayá G, et al. DNN-based performance measures for predicting error rates in automatic speech recognition and optimizing hearing aid parameters. Speech Communication, 2019, 106(6):44-56.

[16] Ri H C. A usage of the syllable unit based on morphological statistics in Korean large vocabulary continuous speech recognition system. International Journal of Speech Technology, 2019, 22(4):971-977.

[17] Cui X, Zhang W, Finkler U, et al. Distributed training of deep neural network acoustic models for automatic speech recognition: A comparison of current training strategies. IEEE Signal Processing Magazine, 2020, 37(3):39-49.

[18] Li D, Zhou Y, Wang Z, et al. Exploiting the potentialities of features for speech emotion recognition. Information Sciences, 2021, 548(6):328-343.

[19] Hülsmeier D, Schdler M R, Kollmeier B. DARF: A data-reduced FADE version for simulations of speech recognition thresholds with real hearing aids. Hearing Research, 2021, 404(2):108-117.

[20] Jermsittiparsert K, Abdurrahman A, Siriattakul P, et al. Pattern recognition and features selection for speech emotion recognition model using deep learning. International Journal of Speech Technology, 2020, 23(4):1-8.

[21] Kawase T, Okamoto M, Fukutomi T, et al. Speech enhancement parameter adjustment to maximize accuracy of automatic speech recognition. IEEE Transactions on Consumer Electronics, 2020, 12(99):1-12.