Hindawi
Scientific Programming
Volume 2022, Article ID 5617400, 11 pages
https://doi.org/10.1155/2022/5617400

Research Article

Acquisition of English Corpus Machine Translation Based on Speech Recognition Technology

Chunyan Jing and Guoying Liu

School of Humanities & Social Sciences, Xi'an Polytechnic University, Xi'an 710048, Shaanxi, China

Correspondence should be addressed to Chunyan Jing; jingcy@xpu.edu.cn

Received 22 June 2022; Revised 30 July 2022; Accepted 11 August 2022; Published 21 September 2022

Academic Editor: Juan Vicente Capella Hernandez

In the present information age, with the rapid development of computer software and hardware technology and the mature manufacturing system of English-speaking enterprises, it is no longer impossible to use statistics for machine translation. The level and quality of machine translation are expected to meet human expectations. This explains the meaning and realization value of the acquisition of English corpus machine translation, introduces the basic principles of speech recognition technology, and combines the characteristics of the English language on the basis of the original Chinese speech recognition system, adopts the technical means of speech recognition, and leads to in-depth research. In the new era of machine translation acquisition of English corpus, we use the combination of LabVIEW and MATLAB to complete the collection, editing, feature extraction, and speech recognition of speech signals and use VQ pattern matching technology to realize the English recognition of a large number of short vocabulary and individuals. In the application part, we applied the classic LabVIEW technology to the speech recognition technology, which actually realized the idea of "software instead of hardware" and achieved better translation results. Experiments show that the accuracy rate of English machine translation can be as high as 94% when using speech recognition technology. According to the results, it takes about 0.2 seconds to complete machine translation for a 30-second speech, which is basically okay to achieve the effect of real-time translation.

1. Introduction

Speech recognition technology is a comprehensive subject based on phonetics theory and computer technology. It is a multidisciplinary research field, usually involving linguistics, acoustics, cognition, artificial intelligence, information processing, and many other subjects. When focused on the learner English corpus, it shows that the learner corpus is a computer-processed text database of the language output of foreign language learners. It can discover the rules and characteristics of interlanguage development by means of parts of speech, errors, semantic coding, or syntactic tagging. Since the 1990s, in order to better conduct interlanguage research on corpora, researchers have developed and constructed multiple learner corpora, such as the Chinese Learner's English Corpus. As a borderline model, speech recognition technology needs research results in many fields as support. From the perspective of computers, it is the intelligent interface part of smart instruments; from the perspective of signal processing, it is part of information recognition; in terms of circuits, communication, electronic systems, signals, and systems, it involves the source processing of communication systems and information; from the perspective of automatic control theory, it can be an important part of pattern recognition; in addition, voice recognition needs another assistance, such as psychology and physiology. Speech recognition is a very difficult research topic. From the point of view of pattern recognition only, the speech signal is a transient event signal, and it is also a time-consuming and unstable random process. There are many kinds of internal issues. Therefore, speech recognition is a kind of multidimensional pattern recognition, which is a very challenging subject.

Machine translation, also known as automatic translation, is the use of computers to convert one natural language (source language) into another natural language (target language). It is a branch of computational linguistics, one of the ultimate goals of artificial intelligence and has important scientific research value. At the same time, machine translation has important practical value. With the rapid development of economic globalization and the Internet, machine translation technology plays an increasingly important role in promoting political, economic, and cultural exchanges. Machine translation has developed today and has been widely used, such as various online translation platforms commonly used by humans, retrieval of information between languages, and various computer-embedded translation programs. Automatic translation has come a long way since the beginning and the development process is very difficult. So far, researchers are studying, and it has not developed very smoothly. Although automatic translation has developed to a large extent compared with before, there are not many products that can be brought to the market. Even these machine translation products, which have been widely used by humans, still have a lot of room for improvement in accuracy. Therefore, the study of machine translation knowledge acquisition is a very important prospect.

Regarding machine translation in the context of speech recognition, Ravanelli said that one area that directly benefits from the latest developments in deep learning is Automatic Speech Recognition (ASR). Ravanelli et al. modified one of the most popular RNN models, the gated recurrent unit (GRU) and proposed a simplified architecture that is very effective for ASR. Ravanelli et al. analyzed the role played by the reset gate and suggested replacing the hyperbolic tangent with a modified linear unit activation. However, this change and batch normalization cannot be combined well, and it is not very helpful for the model to learn the long-term dependence relationship and numerical problems are likely to occur [1]. Watanabe et al. said that the traditional automatic speech recognition (ASR) based on hybrid DNN/HMM is a very complex system consisting of various modules such as acoustics, dictionaries, and language models. It also requires language resources, such as pronunciation dictionaries, tokenization, and speech context dependency trees. However, experiments in English (WSJ and CHiME-4) tasks cannot prove that it is superior to most other popular speech recognition models. The character error rate is relatively increased by 5.4–14.6%, and it shows that it is different from the traditional DNN/HMM ASR without language resources. The performance of the system is quite different [2]. In recent decades, Abdelaziz has proposed many methods of embedding audio and video modes to improve the performance of automatic speech recognition in clear and noisy conditions. However, few studies comparing different AV-ASR fusion models can be found in the literature. However, the implementation process of his research method is relatively redundant, and it is troublesome to operate. At the same time, the research in this paper is based on speech recognition to achieve real-time machine translation acquisition of English corpus, but his research method takes too long to respond to speech recognition, so it is not suitable for this research method [3]. However, the above research was stopped at an early stage due to the imperfection of present English recognition technology and the shortcomings of the English corpus.

In order to complete the effective combination of voice recognition and robot control technology, I created a robot voice control system based on previous research and completed the following tasks: (1) Proposed a new feature parameter extraction method. By improving the LPC spectrum to estimate the formant parameters, new speech feature parameters are constructed. The new feature parameters include more comprehensive voice information including vocal tract characteristics and human auditory characteristics, with high accuracy, strong anti-interference ability, and more accurate highlighting of the essential characteristics of the voice signal. (2) Provides an enhanced speech recognition algorithm (TSMS) and an efficient DTW algorithm. The high-performance DTW algorithm significantly reduces computational time in the recognition process, meets the real-time requirements of the speech signal, and improves system performance to some extent. (3) The speech recognition technology is successfully applied to the motion control of the robot. In the Matlab development environment and VC++ interface, write the program code of the robot voice control system. Through the test of speech input, the perfect combination of speech recognition and machine translation acquisition of English corpus is realized. (4) This article is based on the basic principles of speech recognition and deeply optimizes the machine translation extraction method of the English corpus.

2. Speech Recognition Method and English Corpus Machine Translation Extraction Method

2.1. Basic Principles of the Speech Recognition System. There are many design schemes for speech recognition systems for different tasks, but the system structure and model ideas are roughly the same. The speech recognition system is essentially a pattern recognition system, including three basic units: feature extraction, pattern matching, and reference pattern library. Its basic structure is shown in Figure 1.

The preprocessing module processes the input part of the original audio, removes the data and background noise that do not have much impact on the overall experiment, and performs endpoint detection, voice framing, and lays pre-emphasis on the voice signal. During the training period, the tester conducts multiple training speeches, and the system obtains the feature vector parameters after preprocessing and feature extraction and designs or adjusts the reference model library for training speech [4]. The intermediate results are subsequently processed accordingly, and the constraints of the language model, morphology, syntax, and semantic information are adopted to obtain the final recognition result. After waiting for the input voice frequency band to generate electrical signals through the voice input

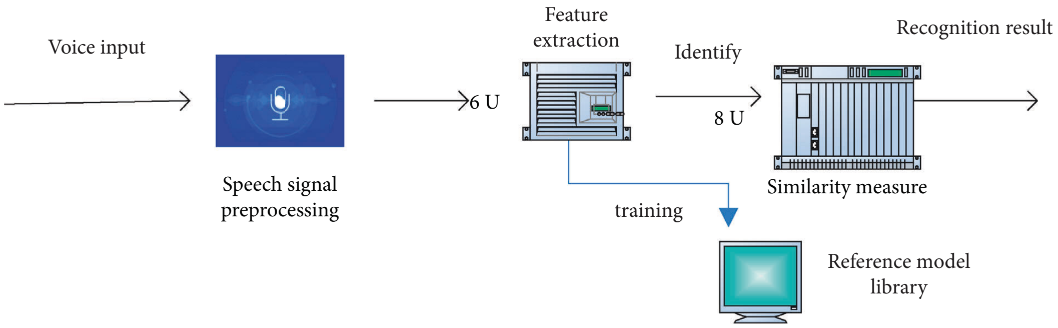

Figure 1: Basic structure diagram of the speech recognition system.

device, add the ports of the recognition system, first do the preliminary work, and then combine various features of the voice to form a voice model, analyze the electrical signal, and form voice recognition on this basis. The template is to be used [5].

2.2. Speech Recognition Technology Mode. The present speech recognition technology mostly adopts the method of pattern matching. As an important branch of pattern recognition, speech recognition and pattern recognition are unified. According to pattern recognition theory, we can compare the speech to be recognized with the existing speech one by one and select the best reference model as the final result of recognition [6], as shown in Figure 2.

As can be seen from Figure 2, the sound is identified by spelling and research on its characteristics, then entered into the database, and finally transferred from the translation library.

In the process of speech recognition, we also use the same method as before, and then we also get a set of speech parameters and then save it as a test template. In order to draw a correct conclusion, we compare the characteristic parameters of the test template and the training template and draw the highest matching rate in the recognition result [7].

The essence of speech recognition is actually the pattern matching process, so we need to train an appropriate template for speech matching. We use a lot of speech data to train this speech model. To a large extent, the nonspecific speech recognition system relies on for the establishment of the speech model library [8]. The process of establishing the voice model library is first select representative speakers, whose voices can evenly represent the general voice distribution. It is best to choose people of different ages and genders as training objects. If we do not select speech objects in this way, even if the speech database is trained, the final speech model will not have a good recognition effect [9].

2.3. Digital Processing of Voice Signals. Under normal circumstances, when we store the sound in the computer, we need to do the digitization of the analog voice, and the digitization of the analog voice signal is divided into two steps: sampling and quantization.

As can be seen from Figure 3, the digital processing part starts with an analog signal, then samples to generate a discrete analog signal, and finally brightens it to achieve the effect of the final form of the digital signal. Sampling is the process of outputting analog signals at equal intervals in the time domain to receive a series of analog audio and convert it to digital audio [10]. In other words, sampling is the process of discretizing the continuous speech signal in a sequence of samples over time (as shown in Figure 3).

$$F(m) = f_b(mT) - \infty < m < +\infty. \tag{1}$$

Among them, m is an integer; T is the sampling period of the voice signal; $H_c = 1/T$ is the sampling frequency of the voice signal.

According to the content of the sampling theorem, if the bandwidth of the frequency spectrum of the speech signal $f_a(s)$ is limited, that is to say

$$f_b(kv) = 0, v > 2\pi K_b. \tag{2}$$

When the sampling frequency is greater than twice the bandwidth of the signal, that is

$$H_t = \frac{1}{T} > 2H_c. \tag{3}$$

The sampling information will not be lost, and the waveform of the original speech signal can be accurately reconstructed from $f(s)$, that is, $f_b(s)$ can uniquely reconstruct the original signal from the sample sequence as follows:

$$f_b(s) = \sum_{n=-\infty}^{+\infty} f_b(mT)\sin\left[\frac{\pi}{T}\left(s - \frac{m}{T}\right)\right]. \tag{4}$$

Among them, $H_t = 2H_c$ is the Nyquist frequency, and the sampling frequency is selected as 8 kHz in this paper. Quantization is to discretize the waveform amplitude value that is discrete in time but still continuous in amplitude. Here, the quantization option is 16 bits. In addition, the channel parameters must also be taken into account. This article uses monophonic parameters. Mono is a relatively original form of sound reproduction, which is more common in early sound cards. When playing back mono information through two speakers, we can clearly feel that the sound is delivered to our ears from the middle of the two speakers.

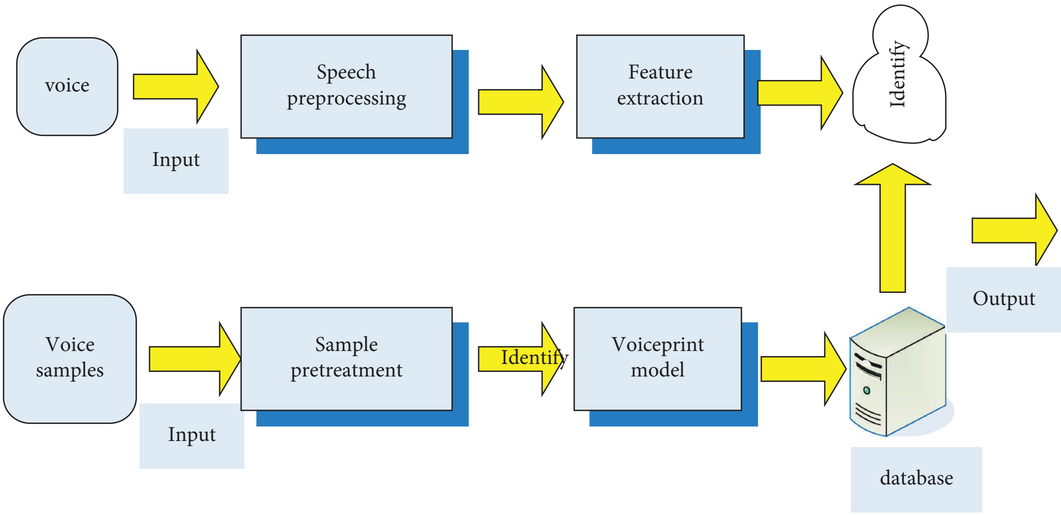

Figure 2: Speech recognition model diagram.

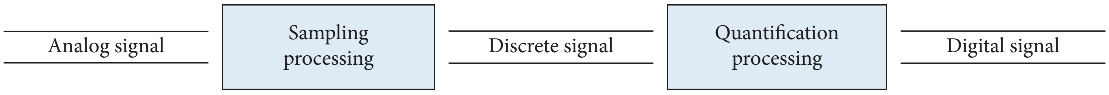

Figure 3: Digital processing process diagram of the speech signal.

2.4. Collection of English Corpus. Along with the mastery and understanding of corpus technology, in order to provide corpus more accurately and in a more standardized way, there are also some small professional corpora. When researchers need to conduct research in historical linguistics, forensic linguistics, language acquisition, and others, the super-large corpus covering a wide range is not suitable. At this time, a special corpus needs to be established to conduct research in a special field. Therefore, while the expanded corpus becomes the trend, the dedicated corpus will also become the future development direction. There are two difficult problems in the construction of English corpus: one is the way to obtain English corpus; the other is how to align the acquired English corpus [11]. Due to the continuous development of the Internet field, bilingual resources are also abundant. When collecting corpus, there are two factors that need to be considered: one is the quality of the plain corpus, and the second is the application range of the parallel corpus. The quality of the corpus itself refers to the fluency of the language, the number of words in each translation sentence pair, and the number of occurrences of each translation unit. The collection of English corpora should be classified according to the application field, article genre, and creation time [12]. For different application goals, different training corpora should be used. If the application goal of the translation system is to face the news domain and collect the English corpus of the news domain as a training corpus, then the accuracy and recall rate of the translation system will be significantly improved [11]. The creation of an English corpus is much more complicated than the creation of other monolingual corpora. When creating an English corpus, it is necessary to perform alignment work, that is, to achieve the sentence-level alignment relationship between the original text and the target text [13]. There are also many ways to store the English body. The first way is to store the source language and the target language in a file at the same time, in the order of the word pairs. Each word pair contains a source language sentence and a sentence target sentence; The two storage methods are to store the source language and target language in two different files [14]. Sometimes we need to label the English corpus, such as part-of-speech labeling and syntactic component labeling. If the syntactic analysis or lexical analysis of the English corpus is carried out, it will bring great help to the research work. The labeling of the English corpus requires the use of different lexical and syntactic labeling tools for the two different language corpora. For example, Chinese and English have different grammatical rules and language habits, so the tools used are also different [15].

Once you have gathered the right body, the next step is to organize and process the body, including word segmentation, lowercase, and root reduction. Sorting and processing tasks must be combined with the characteristics of the body itself. Organize and store in accordance with certain storage formats and specifications [16]. Since the original corpus may come from different collectors, consistency is difficult to guarantee. Including inconsistent storage methods; inconsistent alignment units; inconsistent genres, fields, and creative periods, containing a lot of noise information; inconsistent article layout formats; and duplicate corpus. The goals of organizing the original corpus are (1) the content is consistent and the format is uniform; (2) the alignment unit has basic mark information; (3) noise interference information is deleted. After sorting, the body must be processed in one step, such as formatting and tagging sentences. For different application fields and purposes, different processing methods and strategies are adopted. Since there are still some problems in the automatic sorting, it is necessary to carry out appropriate manual proofreading in the end [17].

2.5. Model Construction. After the extraction of the blocks is completed, the same blocks are merged first, and their frequencies are added up. The same block means that the source phrase and the target phrase are the same, and the alignment is the same. Then use the phrase translation model of the frequency calculation block and smooth it.

$$p(e/c) = \frac{\text{count}(c,e)}{\sum_{e'}\text{count}(c,e') + d},$$
$$p(c/e) = \frac{\text{count}(c,e)}{\sum_{e'}\text{count}(c,e') + d}. \tag{5}$$

Count (c, e) represents the number of occurrences of the block, and p (e/c) is the smoothing parameter.

For each block, define the vocabulary translation model as follows:

$$p_w\left(\frac{e}{c}\right) = p_w\left(\frac{e}{a,c}\right) = \prod_{i=1}^{c}\frac{1}{(i,j)\in a}\sum w\left(\frac{e_i}{e_j}\right),$$
$$p_w\left(\frac{c}{e}\right) = p_w\left(\frac{c}{a,e}\right) = \prod_{j=1}^{c}\frac{1}{(i,j)\in a}\sum w\left(\frac{e_j}{e_i}\right), \tag{6}$$

where, $c(e_j, e_i)$ represents the mapping between the words contained in the block and a is the alignment in the block (c, e).

$$h_{10}(C,A,E) = sim(T_1, A),$$
$$h_{11}(C,A,E) = sim(T_1, A). \tag{7}$$

Therefore, the new word alignment model is

$$\Pr\left(\frac{A}{C,E}\right) = \frac{\exp\left[\sum_m^{11}\lambda_m h_m(C,A,E)\right]}{\sum_{A'}\exp\left[\sum_m^{11}\lambda_m h_m(C,A,E)\right]}. \tag{8}$$

The corresponding decision rules are

$$a = \arg_a \max\left\{\sum_{m=1}^{M}\lambda_m h_m(a,e,f,v)\right\}. \tag{9}$$

Correspondingly, in the speech recognition search algorithm, the scoring function is changed to

$$\text{score}(A) = \sum_{i=1}^{11}\lambda_i h_i. \tag{10}$$

The parameter training and weight adjustment of the model are the same as the previous model.

Energy frequency value method: The energy frequency value method is the product of the short-term energy and the short-term zero transit rate. According to the characteristics of the English language, the initial zero transit rate is relatively high and the final energy is relatively high. The energy frequency method can take into account these two characteristics, making it have a higher resolution ability, and the use of energy frequency can improve the adaptability of the system [18].

The calculation steps of the energy frequency method used in endpoint detection are as follows:

(1) Set a relative threshold value, denoted as T;

(2) calculate their respective short-term energy and average zero-crossing rate according to the data obtained by subframes, and then multiply the calculated data two by two to obtain their respective energy frequency values, which are arranged in a sequence, which can be expressed as FE(0), FE(1), ...;

(3) perform median filtering on the energy frequency value sequence obtained in the first step and obtain a new sequence: fe(0), fe(1), ...;

(4) take any energy frequency value fe(s) at a time s and find the maximum energy frequency value fe($s + i$) in a certain range, namely

$$fe(s) < fe(s+1) < \ldots < fe(s+i)fe(s+i) < fe(s+i+1). \tag{11}$$

(5) Calculate the ratio of the energy frequency value at time $t+i$ to time t, namely

$$k = \frac{fe(s+i)}{fe(s)}. \tag{12}$$

(6) Analyze and judge the above results.

If $s < T$, then time t is not the starting point, assign $s+j+1$ to t and continue following the steps from (4); if $s > T$, time s can be determined to be the starting point of the speech; the endpoint is also found by following this method.

3. English Corpus Machine Translation Platform

3.1. System Design Requirements

(1) Collect voice signals from the sound card that comes with the computer.

(2) Denoising the speech signal.

(3) Use the combination of LabVIEW and MATLAB to complete the collection, processing, feature extraction, and recognition of speech signals.

(4) Realize English speech recognition technology for small vocabulary and isolated words.

3.2. Overall System Design Scheme. Using the classic virtual instrument software LabVIEW, the virtual instrument technology is applied to speech recognition technology, and the idea of "software instead of hardware" is truly realized. Among them, the development platform uses LabVIEW2014, the collection of voice signals uses the computer's own sound card to complete this work, and the processing of voice signals uses Matlab7.0 with powerful data processing capabilities [19].

The purpose of applying virtual instruments to speech recognition systems is that this method can make full use of its flexible graphical programming and has the advantages of strong practicality and low failure rate [20]. In the process of programming, it can simply realize the voice collection, replace the hardware with software, which is simple and easy to understand, and can be updated and upgraded continuously with the development of computer software and hardware and virtual instrument technology, and the cost is relatively low. Therefore, it has a certain practical value and is worthy of research and promotion. For voice extraction, use the audio recorder that comes with Windows to record the words spoken by 20 people, 5 times for each person, a total of 100 times. Choose the clearer 20 voices as reference templates. The experimental feature parameters adopt the extracted MFCC + formant parameters to form a feature vector. In the same environment, on the same dataset, record the time for similarity matching of time series with different lengths and compare the experimental results of different sampling points to compare multiple sets of experiments to analyze the time series of different lengths. Run results on the dataset [21].

Because Matlab has powerful data processing capabilities, it has chosen to use Matlab to realize the design of speech recognition algorithms. Through the LabVIEW platform to manage and call Matlab, the combination of the two realizes the design of the speech recognition system [22]. The speech recognition system first inputs the speech to the computer through the sound card for signal collection and storage and then performs preprocessing, feature extraction, recognition algorithms, and other operations, and finally can get the recognition result [23].

This link mainly introduces the generation of the voice signal and the process of voice signal digitization. In order to ensure that high-quality speech signals are obtained after digitization, filtering must be performed before the original speech signals are digitized. Normally, filtering and digital processing are integrated into one module, so there is no need to design a separate filtering module. Design idea: through the "configure sound input" control, set the computer's own sound card for sound collection parameter settings and then use the "read voice input" control to read the sound into the computer through the microphone and then display the waveform and data. Finally, Save the input sound in a path in.wav format through the "write sound file" and "write and open sound file" controls, and finally use the "sound input clear" control and "close sound file" control to achieve voice signal collection and storage [24]. The program block diagram is shown in Figure 4, and its parameter settings, sampling frequency is 22050, sampling channel is 2, the number of bits per sampling is 16, the number of samples per channel is 5000, and the sampling mode is in continuous sampling.

Due to the limitation of the probability value in GIZA++ results, we further use information entropy to obtain the correct results more effectively. According to the maximum probability value method, we will get the wrong results, but in our method, the threshold (the difference with the maximum probability) is 0.1.

The overall flow of the algorithm is shown in Figure 5.

4. Experimental Results and Discussion

The characteristics of the spectrogram are just like the fingerprint of a person. We call it "voiceprint", which varies from person to person. The voice prints of different speakers are different. The voice parameters can be determined from the characteristics of the spectrogram. Here, the formant is used as the voice parameter to express the voice characteristics of different speakers when they speak the same voice. Tables 1 and 2 give the pitch frequency and formant values extracted from the spectrogram.

From the experimental data given in Table 2, it appears that when the same speaker speaks different words, the fundamental frequency is essentially the same, but the values of the first, second, and third formations are quite different. At the same time, it can be seen that different speakers have different base frequencies when uttering the same word, and the values of the first, second, and third formants are also different. The third formant has a big difference, probably the value is higher for girls. It is possible to obtain new voice feature parameters based on voice recognition that can accurately extract voice features.

After preprocessing, the statistical information of the corpus, development set, and test set is shown in Table 3.

Regarding the recorded speech doped with noise, this article uses the frequency analysis of the speech signal to obtain the spectral entropy value of the speech signal and the noise signal and uses the basic spectral subtraction method to eliminate the noise to obtain a relatively pure speech. The results of each step of the entire denoising process and the analysis process are given below. First, the voice waveform of the recorded experimental voice 'hello' and the waveform of pure Gaussian white noise are given, as shown in Figures 6 and 7:

Then, use the spectrum of the 'hello' speech signal and the Gaussian white noise signal to obtain the power spectrum and construct the spectral entropy function respectively, where the speech signal is represented by S, and the Gaussian white noise is represented by N, and the result shown in Figure 8 can be obtained.

As shown in Figure 8, it can be concluded that since the "hello" voice signal is only concentrated in a few frequency bands, the spectral entropy value of the "hello" voice signal is distributed below 7, while the noise signal is distributed throughout the entire voice spectrum. Range, its spectral entropy value is larger than that of a pure speech signal, is about 8.6.

Take the number of sampling points as 480 points, and experiment with the traditional DTW algorithm and the improved DTW high-efficiency algorithm. The program is run 3 times each time, and the total time consumed is calculated for comparative analysis.

From the experimental data in Table 4, it can be seen that compared with the traditional DTW algorithm and the

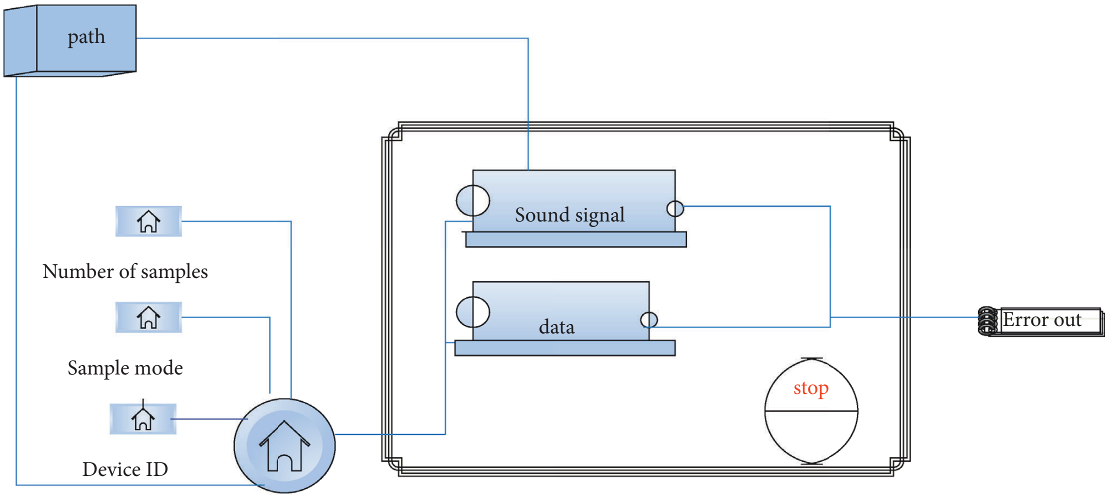

FIGURE 4: The program block diagram of language signal acquisition and storage.

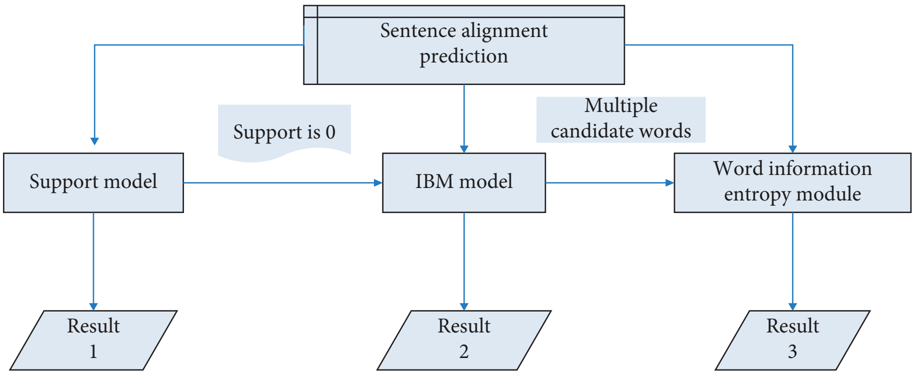

FIGURE 5: Hybrid model word alignment process diagram.

improved DTW high-efficiency algorithm, the time consumed by the improved algorithm has been increased from the results of each run; from the perspective of the total time consumption, the improved DTW when compared with the traditional algorithm, the efficient algorithm has increased by 2.28 s, that is, an increase of 2.19%; in addition, from the point of view of the recognition rate, the two are similar, almost close to the same. In other words, the improved DTW high-efficiency algorithm saves computing time while maintaining a high recognition rate.

Take the number of sampling points as 720 points and experiment with the traditional DTW algorithm and the improved DTW high-efficiency algorithm. Each time the program is run 3 times, the total time consumed is calculated for comparative analysis. And record the recognition rate of the two algorithms.

From the experimental data in Table 5, it can be seen that compared with the traditional DTW algorithm and the improved DTW high-efficiency algorithm, the time consumed by the improved algorithm has been increased from the results of each run; from the perspective of the total time consumption, the improved DTW when compared with the traditional algorithm, the efficient algorithm is increased by 7.57 s, which is an increase of 3.1%; for the recognition rate, the two are still close to the same.

In this experiment, 500 English phrases are randomly selected by the 10,000-sentence teaching body as the test to conduct an English phrase alignment experiment. When $K = 1$, the number of English phrases that can be extracted is 406, of which the correct number is 312; when $K = 2$, the number of English phrases that can be extracted is 368, of which the correct number is when $K = 3$, the number of English phrases that can be extracted is 321, of which the correct number is 287. According to the different values of K, the experimental results we get are shown in Figure 9:

It can be seen from the figure that when the value is larger, the accuracy rate will increase because the standard for defining alignment is more stringent, but the recall rate will decrease because the total number of phrases that the system can get has decreased.

Table 6 shows the experimental results given by each system. Among them, IBM4E- > C represents the word alignment of the IBM model from English to Chinese. Similarly, IBM4C- > E represents the word alignment of the IBM model from Chinese to English and intersection, union, and refined respectively indicate two alignment directions.

Table 1: Comparison of the parameters of the sound “thank you” pronounced by the same speaker.

Same speaker	Pitch frequency (Hz)	First formant (Hz)	Second formant (Hz)	Third formant (Hz)
‘Thank you’	205.85	415.62	2147.36	3124.98
‘Hello there’	197.35	605.78	1633.25	2749.15

Table 2: Comparison of parameters of the same word ‘thank you’ sent by different speakers.

Different speakers	Pitch frequency (Hz)	First formant (Hz)	Second formant (Hz)	Third formant (Hz)
Speaker A (male)	135.45	408.25	2025.36	2514.36
Speaker B (man)	139.56	421.08	1986.15	2459.36
Speaker C (female)	201.62	426.61	2238.92	3012.26

Table 3: Statistics of word alignment experiment corpus.

	Training corpus	Concurrent set	Test set
Number of sentences	333901	502	504
Word count	5582631	9338	13902
Number of links	74264	6389	14046

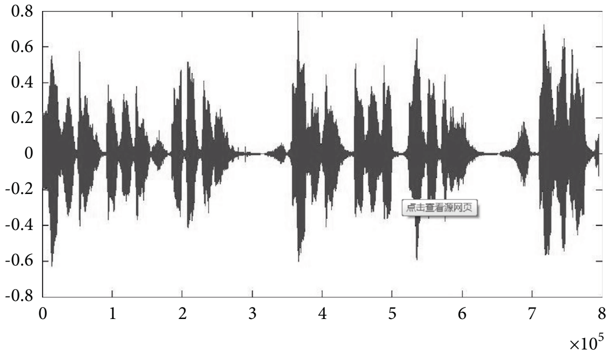

Figure 6: “Hello” voice waveform.

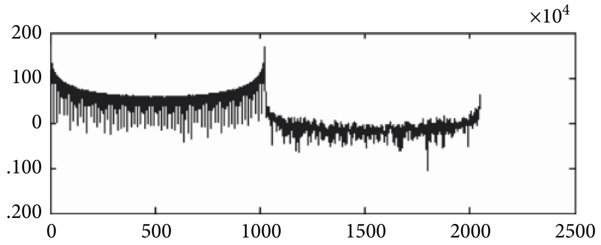

Figure 7: Pure white Gaussian noise speech waveform.

Find the intersection, union, and word alignment obtained by the grow-diag-final rule. ITG-M means that only constraints are used in word alignment. ITG-S means to combine ITG and syntax trees.

By comparing the results, it can be found that word alignment with syntactic constraints has achieved better alignment results than the IBM-4 model. Compared to the more sophisticated IBM-4, the ITG-S improves by about 5%. Except for the two weakly constrained syntactic knowledge, the features adopted by ITG-M and ITG-S are similar to the submodels in the IBM model. Therefore, it is effective to combine syntactic knowledge in word alignment.

For the systems ITG-M and ITG-S, ITG-S has achieved a better AER value. By analyzing the specific word alignment content of the experiment, it is found that due to the constraints of the syntax tree, the word alignment links in

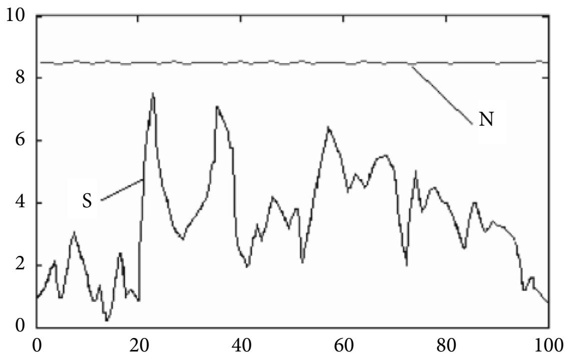

Figure 8: Spectral entropy of S and N.

Table 4: Operation time and recognition rate when the number of sampling points is 480 (%).

Algorithm	First run time	Second run time	Third run time	Total time	Recognition rate
Traditional DTW algorithm	34.83	34.69	34.77	104.29	98.83
DTW efficient algorithm	34.16	33.96	33.89	102.01	98.79
TSMS	34.35	34.20	34.26	102.81	98.82

Table 5: Operation time and recognition rate when the number of sampling points is 720 (%).

Algorithm	First run time	Second run time	Third run time	Total time	Recognition rate
Traditional DTW algorithm	138.54	140.53	142.38	421.45	98.52
DTW efficient algorithm	127.67	131.39	131.84	390.9	98.66
TSMS	132.20	134.35	134.26	401.81	98.61

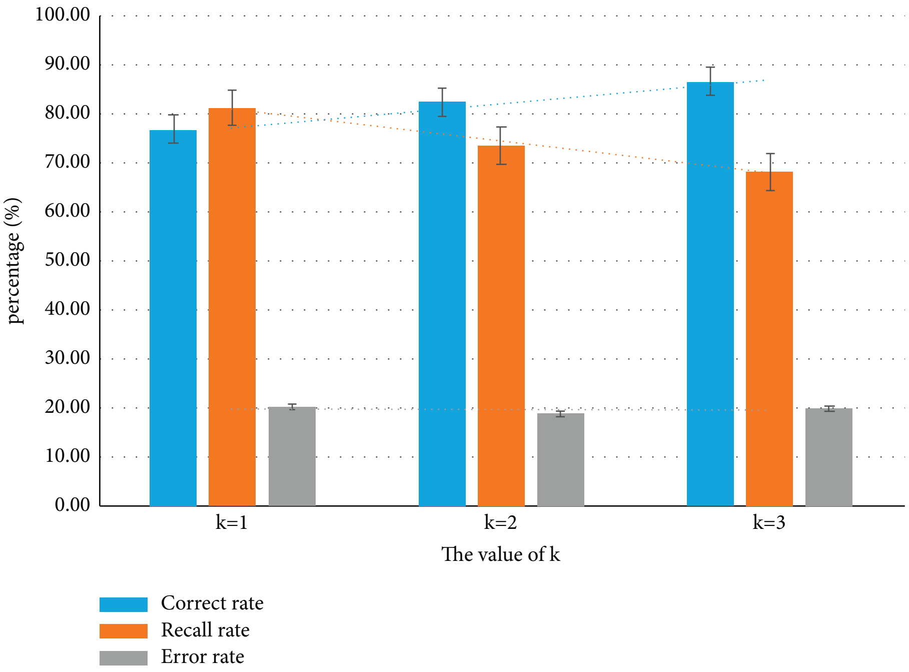

Figure 9: Different experimental results with different values.

Table 6: Comparison of experimental results of the word alignment system.

System	Accuracy	Recall rate	AER
IBM4E- > C	0.8164	0.6091	0.3013
IBM4E- > E	0.7856	0.5557	0.3478
IBM4 intersection	0.9549	0.4912	0.3497
IBM4 union	0.7181	0.6737	0.3045
IBM4 refined	0.8694	0.6002	0.2885
ITG-M	0.8305	0.6296	0.2826
ITG-S	0.8335	0.6408	0.2743

ITG-S are more biased. Because of clustering together, the number of links with a large span is relatively small, which is conducive to obtaining a higher accuracy of word alignment.

5. Conclusions

On the basis of summarizing predecessors' speech recognition in other languages, this paper analyzes and studies the basic pronunciation and phonetic characteristics of English speech, and systematically studies the related problems of English speech recognition. Speech recognition technology has extremely broad application prospects, coupled with the idea that "software is an instrument", which stimulates further exploration of speech recognition technology. In this article, the combination of LabVIEW and MATLAB is used to complete the collection, processing, feature extraction, and recognition of speech signals. It is the first to collect voice signal data through the sound card that comes with the computer, achieving the function of real-time voice signal collection and storage. Secondly, the speech signal is denoised. There are many kinds of noise in the speech signal. In view of the characteristics of these noises, the wavelet denoising processing method is selected, that is, after the effective speech signal is processed by the wavelet transform, most of the information located in the low-frequency part of the wavelet transform frequency band is larger, and the white noise is mostly located in the high-frequency part of the smaller wavelet transform frequency band. In the process of speech recognition, the speech signal after framing and windowing must be detected by endpoints. A combination of the short-term energy method and short-term average zero-crossing rate method can be used to improve the stability and recognition rate of the system. Speech recognition using the energy frequency method and improved DTW high-efficiency algorithm can save computing time while maintaining a high recognition rate. The average translation accuracy rate can be as high as 95%, and the average response time is about 0.5 seconds, which can fully meet the requirements of daily English corpus machine translation.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare no potential conflicts of interest in this study.

Acknowledgments

This work was supported by Foreign Language Research Project with Social Sciences Association of Shaanxi Province: A Study on English Translation and Communication of Cultural Relics of Han Dynasty Mausoleum in Xi'an from the Perspective of Compensation Translation (2021ND0609).

References

[1] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light gated recurrent units for speech recognition," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 92–102, 2018.

[2] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[3] H. A. Abdelaziz, "Comparing fusion models for DNN-based audiovisual continuous speech recognition," *IEEE/ACM Transactions on Audio Speech & Language Processing*, vol. 26, no. 3, pp. 475–484, 2017.

[4] R. Randi and I. Nancy, "The American national corpus," *Journal of English Linguistics*, vol. 32, no. 2, pp. 105–113, 2016.

[5] F. G. Chong, G. Friedland, A. Janin, N. Morgan, and C. Oei, "Opportunities and challenges of parallelizing speech recognition," *Diabetes Technology & Therapeutics*, vol. 18, no. 6, p. 2, 2016.

[6] T. Y. Ahn and S. M. Lee, "User experience of a mobile speaking application with automatic speech recognition for EFL learning," *British Journal of Educational Technology*, vol. 47, no. 4, pp. 778–786, 2016.

[7] R. Yazdani, A. Segura, J. M. Arnau, and A. Gonzalez, "Low-power automatic speech recognition through a mobile GPU and a viterbi accelerator," *IEEE Micro*, vol. 37, no. 1, pp. 22–29, 2017.

[8] S. Gordon-Salant and S. S. Cole, "Effects of age and working memory capacity on speech recognition performance in noise among listeners with normal hearing," *Ear and Hearing*, vol. 37, no. 5, pp. 593–602, 2016.

[9] O. T. Grozdi and S. T. Jovii, "Whispered speech recognition using deep denoising autoencoder," *Engineering Applications of Artificial Intelligence*, vol. 59, no. 12, pp. 2313–2322, 2017.

[10] B. Ren, L. Wang, L. Lu, Y. Ueda, and A. Kai, "Combination of bottleneck feature extraction and dereverberation for distant-talking speech recognition," *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5093–5108, 2016.

[11] M. Kim, B. Cao, T. Mau, and J. Wang, "Speaker-independent silent speech recognition from flesh-point Articulatory movements using an LSTM neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2323–2336, 2017.

[12] P. Sharma, A. Vinayak, and A. K. Sao, "Deep-sparse-representation-based features for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2162–2175, 2017.

[13] S. S. Wang, P. Lin, Y. Tsao, J. W. Hung, and B. Su, "Suppression by selecting wavelets for feature compression in distributed speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 564–579, 2018.

[14] A. Biswas, P. K. Sahu, and M. Chandra, "Multiple cameras audio visual speech recognition using active appearance

model visual features in car environment," *International Journal of Speech Technology*, vol. 19, no. 1, pp. 159–171, 2016.

[15] X. Ji, J. Pan, and Y. Yan, "Agglutinative language speech recognition using automatic allophone deriving," *Chinese Journal of Electronics*, vol. 25, no. 2, pp. 328–333, 2016.

[16] A. Xing, Q. Zhao, and Y. Yan, "Speeding up deep neural networks in speech recognition with piecewise quantized sigmoidal activation function," *IEICE - Transactions on Info and Systems*, vol. E99.D, no. 10, pp. 2558–2561, 2016.

[17] S. Mirsamadi and J. H. L. Hansen, "A generalized nonnegative tensor factorization approach for distant speech recognition with distributed microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1721–1731, 2016.

[18] D. Nguyen, X. Xiong, E. S. Chng, and H. Li, "Feature adaptation using linear spectro-temporal transform for robust speech recognition," *IEEE/ACM Transactions on Audio Speech & Language Processing*, vol. 24, no. 6, pp. 1006–1019, 2016.

[19] F. Klubika, A. Toral, and V. M. Sánchez-Cartagena, "Quantitative fine-grained human evaluation of machine translation systems: a case study on English to Croatian," *Machine Translation*, vol. 32, no. 1, pp. 1–21, 2018.

[20] M. Chen, J. Pan, Q. Zhao, and Y. Yan, "Multi-task learning in deep neural networks for Mandarin-English code-mixing speech recognition," *IEICE - Transactions on Info and Systems*, vol. E99.D, no. 10, pp. 2554–2557, 2016.

[21] I. H. Y. Ng, K. Y. S. Lee, J. H. S. Lam, C. A. van Hasselt, and M. C. F. Tong, "An application of item response theory and the rasch model in speech recognition test materials," *American Journal of Audiology*, vol. 25, no. 2, pp. 142–152, 2016.

[22] S. Wu, D. Zhang, Z. Zhang, N. Yang, M. Li, and M. Zhou, "Dependency-to-Dependency neural machine translation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2132–2141, 2018.

[23] R. Miyahara, K. Oosugi, and A. Sugiyama, "A hearing device with an adaptive noise canceller for noise-robust voice input," *IEEE Transactions on Consumer Electronics*, vol. 65, no. 4, pp. 444–453, 2019.

[24] H. Tobias, M. Farah, and C. Enrico, "Efficiency and safety of speech recognition for documentation in the electronic health record[J]," *Journal of the American Medical Informatics Association*, no. 6, pp. 1127–1133, 2017.