Received: 23 December 2024 | Revised: 22 January 2025 | Accepted: 2 February 2025 | Published online: 6 February 2025

International Journal of Advance in Applied Science Research, Volume 4, Issue 2, 2025
https://h-tsp.com/

# Modeling of Speech Recognition Based on Deep Learning

**Min Zhang***

School of Computer Science, Xianyang Normal University, Xianyang, Shaanxi 712000, China
**Author to whom correspondence should be addressed.*

**Abstract:** ***As technology continues to advance, the application of speech recognition is becoming increasingly pervasive, and the significance of intelligent speech recognition cannot be overstated. This article delves into the intricate workings and classifications of speech recognition systems, meticulously outlining the process of designing the system's development environment and framework. It meticulously charts the course from the collection of speech datasets to the preprocessing of speech data, and then progresses to the crucial stages of feature extraction and the construction of both acoustic and language models tailored for deep learning-based Chinese speech recognition. This comprehensive study not only enables the system to record speech autonomously or upload pre-recorded speech to a server for Chinese recognition but also boasts the capability to translate the recognized Chinese speech into English. This functionality underscores the study's potential to pave the way for further in-depth exploration and advancements in the realm of speech recognition, establishing a solid foundation for future research endeavors.***

**Keywords:** Deep learning; Speech recognition; Feature extraction; DFSMN model.

## 1. Introduction

Speech recognition technology, also known as Automatic Speech Recognition (ASR), has gradually become closely related to people's lives. The goal of speech recognition is to convert people's voices into binary language that computers can understand and process accordingly. Like text classification and machine translation, speech recognition is a subfield of natural language processing (NLP) in artificial intelligence. In the highly popular era of artificial intelligence, from Siri to Xiaodu, from Xiaobing to Xiaona, and then to Xiaoai Tongxue, these intelligent voice assistants are integrating into people's lives. The application fields of speech recognition technology are very wide, including smart homes, mobile devices, intelligent customer service, in car systems, intelligent healthcare, industrial control, intelligent toys, etc. Its core is to interact with machines through voice and enable them to complete related tasks.

In recent years, advancements in various fields have led to significant breakthroughs in technology and research. Long et al. [1] presented a study in September 2024 at the IEEE 7th International Conference on Information Systems and Computer Aided Education (ICISCAE), enhancing educational content matching using Transformer models and InfoNCE loss. Their work aimed to improve the accuracy and efficiency of content matching in educational platforms. Meanwhile, Huang et al. [2] conducted research on a multi-agency collaboration medical images analysis and classification system based on federated learning, discussed at the 2024 International Conference on Biomedicine and Intelligent Technology. This study explored the potential of federated learning in facilitating secure and efficient collaboration among medical institutions. Ukey et al. [3] published an article in the World Wide Web journal in 2023, introducing an efficient continuous kNN join algorithm for dynamic high-dimensional data. Their work addressed the challenges associated with processing and querying large-scale, high-dimensional datasets. Chen et al. [4], at the 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23), presented a study on channel-aware 5G RAN slicing with customizable schedulers, contributing to the advancement of network slicing technologies in 5G networks. Peng et al. [5] proposed a dual-augmentor framework for domain generalization in 3D human pose estimation at the IEEE/CVF Conference on Computer Vision and Pattern Recognition in 2024. Their framework aimed to improve the robustness and generalization ability of 3D human pose estimation models. Yan et al. [6], in a study published in Sustainability in 2024, examined the impact of CEO power on green innovation and organizational performance in manufacturing firms, adopting a mediational approach. Ren et al. [7], in the Alexandria Engineering Journal in 2025, presented an IoT-based system for 3D pose estimation and motion optimization for athletes, leveraging C3D and OpenPose technologies. Their work demonstrated the potential of IoT in sports training and performance improvement. Fan et al. [8], in an arXiv preprint, conducted research on the online update method for retrieval-augmented generation (RAG) models with incremental learning.

## 2. Speech Recognition System

### 2.1 Working mechanism of speech recognition system

The task of speech recognition is to convert speech sequences into text sequences. There are two ways of speech recognition conversion, one is to directly convert speech into text, and the other is to first convert speech into phonemes (or pinyin) and then convert it into text. Due to the large number of homophones in Chinese, the same sound can represent different characters, which greatly tests the contextual context of the input speech. While training the text results, it is also necessary to consider contextual coherence, which greatly increases the difficulty of training and disperses the training effort, ultimately resulting in low recognition accuracy. Therefore, the feasibility of the first method is not high. Compared to the first recognition method, the second method can allocate training content reasonably, that is, only training the conversion of speech sequences into phonemes and only training the conversion of phonemes into text. This can greatly improve the speech recognition rate. The speech recognition design in the article will be developed around the second method [2-3].

### 2.2 Classification of Speech Recognition Systems

Speech recognition technology can be divided into three categories based on different application scenarios: restricting the way users speak, limiting the range of words users use, and limiting the system's usage Household objects [4-5].

Limit the way users speak. According to the limitations of speech recognition systems on users' speaking styles, they can be divided into isolated word recognition systems, continuous speech recognition systems, and improvisational oral speech recognition systems. Continuous speech recognition system refers to a recognition system that uses medium to large-scale vocabulary but uses subwords as basic recognition units; Due to the random nature of the input, the speech content is also random, accompanied by many random events such as swallowing, stuttering, repetition, hesitation, coughing, wheezing, etc. These characteristics make improvisational speech recognition challenging. Limit the range of words used by users. The scope of vocabulary can be divided into three types: small vocabulary, medium vocabulary, large vocabulary, and unlimited vocabulary.

### 2.3 Main issues of speech recognition

(1) Human factors:

Different people have different ways of speaking, accents, speed, and volume; The way the same person speaks will also change according to the speaker's emotional changes and physical condition. Angry people speak with a faster pace and higher pitch, while sick people speak with a slower pace and lower volume. Each person's speaking style will change over time, which can lead to a decrease in speech recognition rate.

(2) Environmental noise:

In actual recording scenarios, there are often different environmental noises such as car horns, wind and rain, white noise, and the sound of others speaking. When these noises are recorded into the audio, they have a serious impact on speech recognition, leading to a decrease in recognition rate.

(3) Hardware factors:

Different recording devices have different recording performance and recording methods, and the sampling frequency and speech signal of the collected audio are also different, which affects the correct recognition of speech.

(4) The understanding of semantics by speech recognition systems:

Human speech has different forms such as vocabulary and homophones, and the same pronunciation sometimes corresponds to different vocabulary in different contexts. Therefore, it is necessary to establish rules for understanding semantics.

## 3. Design Of Chinese Speech Recognition Framework Based On Deep Learning

The Chinese speech recognition architecture based on deep learning is divided into two parts: acoustic model training and model utilization. The training of acoustic models first requires collecting a dataset, followed by preprocessing of the dataset, and then extracting feature vectors from the speech data one by one, which are input into the acoustic model. The parameters of the neural network are calculated using the Error Back Propagation algorithm (BP algorithm), and finally an ideal model with less acoustic loss is trained. After training the acoustic model, you can input speech data into the trained acoustic model by yourself. After the acoustic model obtains the pinyin recognition result, it inputs the pinyin recognition result into the language model. The language model calculates the most likely word combination based on simple word frequency statistics, queries the word frequency dictionary, converts pinyin into Chinese characters, and finally outputs the entire text. The architecture of the Chinese speech recognition system is shown in Figure 1.

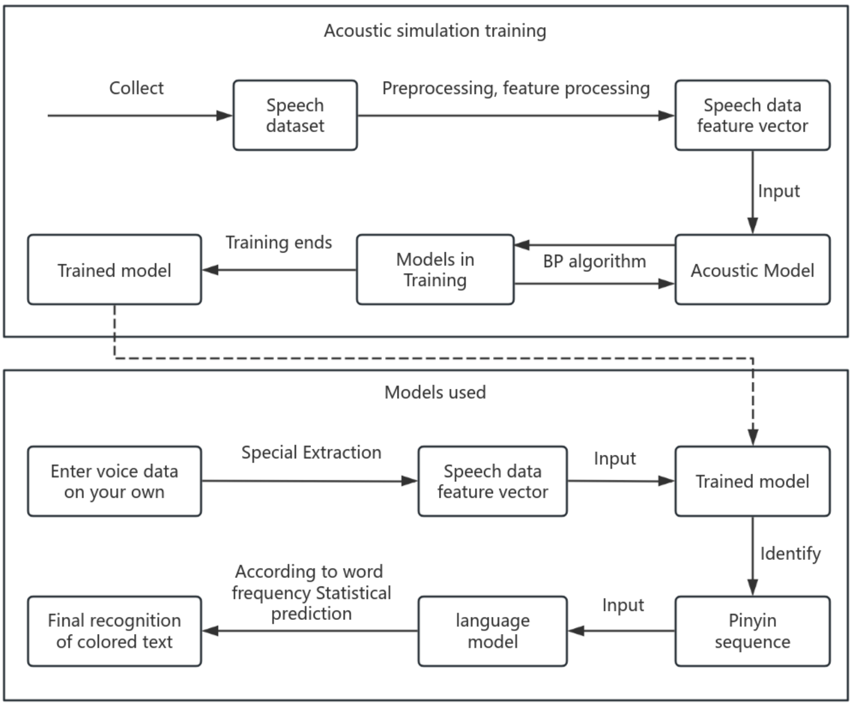

Acoustic simulation training
Collect
Speech dataset
Preprocessing, feature processing
Speech data feature vector
Input
Trained model
Training ends
Models in Training
BP algorithm
Acoustic Model
Models used
Enter voice data on your own
Special Extraction
Speech data feature vector
Input
Trained model
Identify
Final recognition of colored text
According to word frequency Statistical prediction
language model
Input
Pinyin sequence

**Figure 1:** Framework structure of Chinese speech recognition system

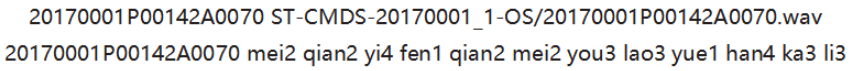
## 4. Collection And Processing Of Voice Data

### 4.1 Collection of Voice Data

The speech mentioned in the article, including speech datasets, device recorded audio, and self uploaded audio, has a sampling rate of 16kHz. Collecting speech datasets is the first step in speech recognition, and many commercial speech recognition datasets are not developed for the public. However, there are still some publicly available high-quality datasets for training. Common Chinese speech datasets include THCHS-30, AISHELL, Magicdata, Primewords Chinese Corpus Set 1, Aidatatang_200zhST CMDS, etc. These six datasets have a total of approximately 1385 hours. Due to limited computer resources, the Chinese speech recognition system chose to use three datasets, THCHS-30, ST-CMDS, and Primewords, to train the acoustic model.

### 4.2 Preprocessing of Voice Data

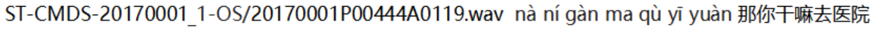
Speech data preprocessing mainly involves dividing the dataset and calibrating label data. While preparing the voice data, it is necessary to calibrate the label data for the voice data. Each training result needs to be compared with the correct result, and the backpropagation algorithm is used to backpropagate the error of the loss function from the output layer to the hidden layer and back to the input layer layer layer by layer. The error is distributed to the units of each layer, thus achieving the goal of updating and solving the weight value W and bias value b. When training the acoustic model with input speech data, simply enter txt text, which includes labels. The directory of a speech file corresponds to its corresponding pinyin sequence and Chinese text sequence. In the data tag, one piece of data references two txt tag texts, so there are a total of 12 txt tag texts. The data tags are shown in Figure 2.



20170001P00142A0070 ST-CMDS-20170001_1-OS/20170001P00142A0070.wav
20170001P00142A0070 mei2 qian2 yi4 fen1 qian2 mei2 you3 lao3 yue1 han4 ka3 li3

**Figure 2:** Sample of Original Label

The method for improving the calibration label data in the article is shown in the specific form of waf_data_cath \ tpinyinlist \ thanzitlist, as shown in Figure 3. You can see that the pinyin characters in the tag have phonetic symbols, but when the acoustic model outputs, it actually generates a series of serial numbers. You can create a JSON dictionary package yourself, which is the model \ model_1anguage \ pinyin-dict.jsn file, responsible for corresponding these serial numbers to a string of letters with numbers at the end, such as na4, ni3, etc., instead of pinyin characters with built-in phonetic symbols.



ST-CMDS-20170001_1-OS/20170001P00444A0119.wav nà ní gàn ma qù yī yuàn 那你干嘛去医院

**Figure 3:** Sample Label in the Text

The text above the voice only represents a label for a voice data, with 9 label files, namely thchs_train.exe of THCHS-30 Thchs_dev.txt, thchs_test. txt, stCMD_train. txt for ST-CMDS Stcmd_dev. txt, stcmd_test. txt, and Prime's prime_train. txt prime_dev.txt、prime_test.txt, Each txt file contains the corresponding number of data labels. The WAV speech dataset file and label text are placed together, and during acoustic model training, the root directory of the dataset, datasets, can be directly introduced.

### 4.3 Speech data feature processing

When training a certain content, it is necessary to extract the features of the content to be trained, and the field of speech recognition is no exception. First, the speech features are extracted, and the neural network will make judgments and classifications based on these characteristics during training. In terms of feature information, Fbanks have more features than MFCC, which includes steps such as Discrete Cosine Transform (DCT) and requires more computation. This is actually a loss of speech information, resulting in a significant loss of sound details. This Chinese speech recognition uses the Fbank feature extraction method.

## 5. Modeling Of Chinese Speech Recognition Based On Deep Learning

### 5.1 Building Acoustic Models

Acoustic models are used in automatic speech recognition systems to represent the relationship between audio signals and phonemes or other language units that make up speech. Modern speech recognition systems use acoustic models and language models to represent the statistical properties of speech. The acoustic model simulates the functional relationship between the processed speech features and speech in language. Then, using a language model, the top-level word sequence corresponding to the given audio clip will be obtained. The establishment of an acoustic model is the most critical part of the entire speech recognition system, and the quality of a speech recognition system largely depends on the quality of the acoustic model in the system. Most modern speech recognition systems run on small pieces of audio, namely frames, with a duration of approximately 10ms per frame. The original audio signal of each frame can be transformed using feature extraction methods such as mel frequency cepstral analysis. The coefficients of this conversion are commonly referred to as Mel Frequency Cepstral Coefficients (MFCC) s and are used as inputs to the acoustic model along with other features. The audio of the acoustic model can use different sampling rates, and the sampling rate used for training the acoustic model should ideally have the same sampling rate and bit recording as the recognized speech audio in order to achieve the best speech recognition performance.

### 5.2 Building Language Models

To build a statistical language model, it is necessary to first collect a sufficiently large amount of word frequency statistical text, including monosyllabic words, disyllabic words, etc. Using probability and statistics methods to construct a language model, inputting a Pinyin sequence, one can obtain the sequence of Chinese characters with the highest probability of occurrence, and then output it as the most reasonable sentence. In statistical language models, the appearance of each word is only considered to be related to the preceding word. Usually, considering the previous word or the first two words, a sufficiently high accuracy can already be obtained, which are called statistical unary language models and statistical binary language models, respectively. In rare cases, ternary, quaternary, and other language models are considered. However, the higher the level of the element, the higher the computational time complexity. When dealing with long pinyin texts, ordinary computer accountants find it very difficult to calculate, resulting in unavoidable time costs. We collected word frequency statistical dictionaries for unary and binary words in this language model, with data volumes of 6880 and 568647, respectively.

Based on Markov chain, achieve the conversion of Pinyin to text. Markov chain is implemented based on dynamic programming algorithm, similar to the algorithm for finding the shortest path. The matching between Chinese characters and Pinyin can be seen as a communication between homophones and Pinyin, with matching done from left to right, as shown in Figure 4.

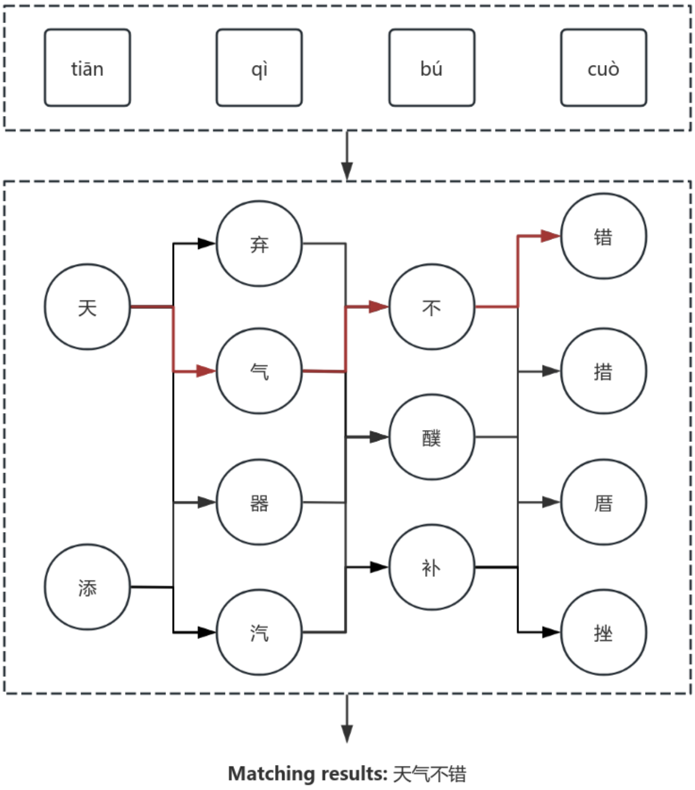


**Figure 4:** Directed Pinyin to Chinese Character Conversion Diagram

### 5.3 Instructions for Using Training Model Code Files

The folder "modelCode" for training models contains the following files:

**cnn_dfsmn_ctc:** Store the trained acoustic model.

**datasets:** Store voice dataset and label files.

**model:** Contains the model_1anguage folder, the acoustic models AcousticVNet py and LanguageVNet py, and the text required for the acoustic and language models in the model_1anguage folder, including the pinyin sequence pinyin-dictjson, the single word frequency statistics text language-word1.txt, the double word frequency statistics text language-word2.txt, and the pinyin dictionary dict. txt.

**plain.py:** Contains functions for drawing time-domain, frequency-domain, and spectrogram graphs.

**train_and_test.py:** Contains functions for training and loading acoustic models. Speech recognition testing can be done by calling the load acoustic model function.

**wav_speech_recorder.py:** Contains functions for recording WAV speech.

## 6. Conclusion

The use of the DFSMN framework in the article, which can model the dependencies between sequences before and after, has improved the recognition rate of the model. The paper first introduces the speech recognition system, explains its working mechanism and classification, and then designs the system development environment and overall framework. Finally, a detailed description of the steps and methods involved in the design process was provided, including the collection of speech datasets, preprocessing of speech data, feature extraction of speech data, construction of acoustic models, and construction of language models. From an application perspective, this application can achieve the function of self recording voice on the web or uploading voice to the server for Chinese recognition, and support the function of translating recognized Chinese into English. From a research perspective, although speech recognition technology involves complex disciplines and techniques, the overall architecture, including dataset collection, feature extraction, acoustic model framework selection, neural network design, and language model establishment, is relatively scientific and can lay the foundation for further in-depth research.

## References

[1] Long, Y., Gu, D., Li, X., Lu, P., & Cao, J. (2024, September). Enhancing Educational Content Matching Using Transformer Models and InfoNCE Loss. In 2024 IEEE 7th International Conference on Information Systems and Computer Aided Education (ICISCAE) (pp. 11-15). IEEE.

[2] Huang, S., Diao, S., Wan, Y., & Song, C. (2024, August). Research on multi-agency collaboration medical images analysis and classification system based on federated learning. In Proceedings of the 2024 International Conference on Biomedicine and Intelligent Technology (pp. 40-44).

[3] Ukey, N., Zhang, G., Yang, Z., Li, B., Li, W., & Zhang, W. (2023). Efficient continuous kNN join over dynamic high-dimensional data. World Wide Web, 26(6), 3759-3794.

[4] Chen, Y., Yao, R., Hassanieh, H., & Mittal, R. (2023). {Channel-Aware} 5g {RAN} slicing with customizable schedulers. In 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23) (pp. 1767-1782).

[5] Peng, Q., Zheng, C., & Chen, C. (2024). A Dual-Augmentor Framework for Domain Generalization in 3D Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2240-2249).

[6] Yan, Q., Yan, J., Zhang, D., Bi, S., Tian, Y., Mubeen, R., & Abbas, J. (2024). Does CEO power affect manufacturing firms' green innovation and organizational performance? A mediational approach. Sustainability, 16(14), 6015.

[7] Ren, F., Ren, C., & Lyu, T. (2025). Iot-based 3d pose estimation and motion optimization for athletes: Application of c3d and openpose. Alexandria Engineering Journal, 115, 210-221.

[8] Fan, Y., Wang, Y., Liu, L., Tang, X., Sun, N., & Yu, Z. (2025). Research on the Online Update Method for Retrieval-Augmented Generation (RAG) Model with Incremental Learning. arXiv preprint arXiv:2501.07063.