

# Research on Modeling of Speech Recognition Based on Deep Learning

Min Zhang

Huizhou University Network and Information Center, Huizhou, Guangdong 516007

**Abstract:** *With the development of technology, the application of speech recognition is becoming increasingly widespread, and intelligent speech recognition has great significance. The article introduces the working mechanism and classification of speech recognition systems, and designs the system development environment and framework. Design the collection of speech datasets, preprocessing of speech data, feature extraction of speech data, construction of acoustic models, and construction of language models for deep learning based Chinese speech recognition. This study is capable of self recording speech or uploading speech to a server for Chinese recognition, and supports the function of translating recognized Chinese into English. This study can lay the foundation for further in-depth research on speech recognition.*

**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 Xiaoi 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. Wang et al. (2022) pioneered spatial immunology by developing a cell atlas of dendritic cell heterogeneity in gastrointestinal cancers, revealing new immune microenvironment dynamics [1]. Concurrently, Li et al. (2025) introduced gamification techniques for smart city data visualization, significantly enhancing public engagement with urban monitoring systems [2], while Wang (2025) optimized urban logistics through Bayesian network reconfiguration algorithms [3]. In clinical research, Li (2025) improved pharmacovigilance in chronic disease trials through machine learning-enhanced adverse event detection [4], complemented by Li et al.'s (2025) work on named entity recognition for real-time smart city data processing [5]. Financial analytics saw progress through Yang's (2025) application of LightGBM in Chinese market predictions [6], paralleled by Song's (2025) AI-driven operational tools for e-commerce platforms [7]. High-performance computing breakthroughs include Wu's (2024) cloud architecture for genetic disease modeling [8] and Chen's (2025) scalable data lakes for autonomous vehicles [9]. Materials science benefited from Lin et al.'s (2025) transfer learning approach to nanophotonic array modeling [10], while Gong et al. (2024) optimized enterprise risk systems via ensemble machine learning [11]. Finally, Bohang et al. (2025) advanced image forensics through active learning-based steganalysis [12], collectively showcasing transformative AI applications across sectors.

## 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.

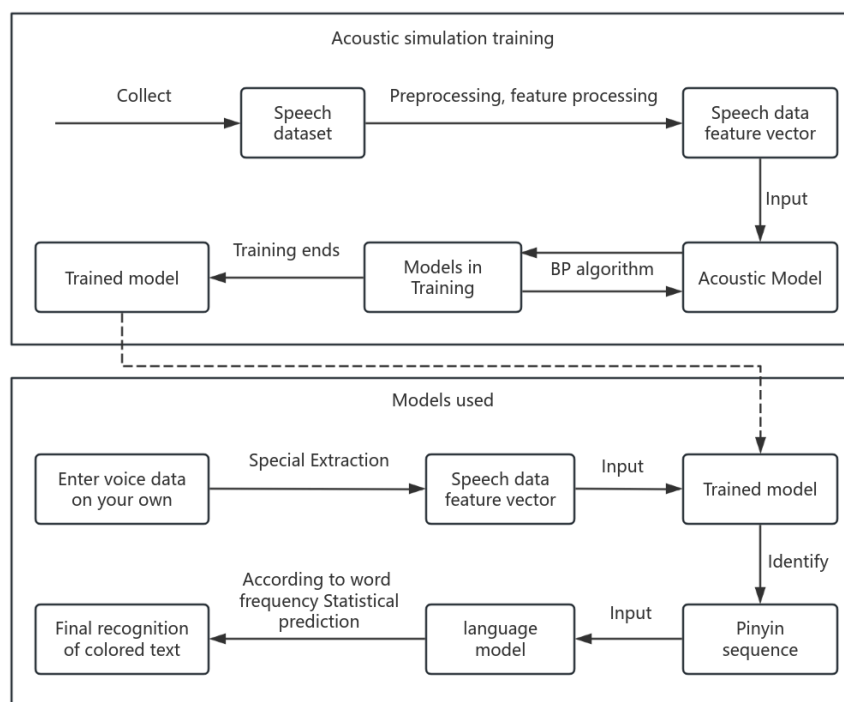
### **(4) The understanding of semantics by speech recognition systems:**

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.



**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\_language \ pinyin-dict.json 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

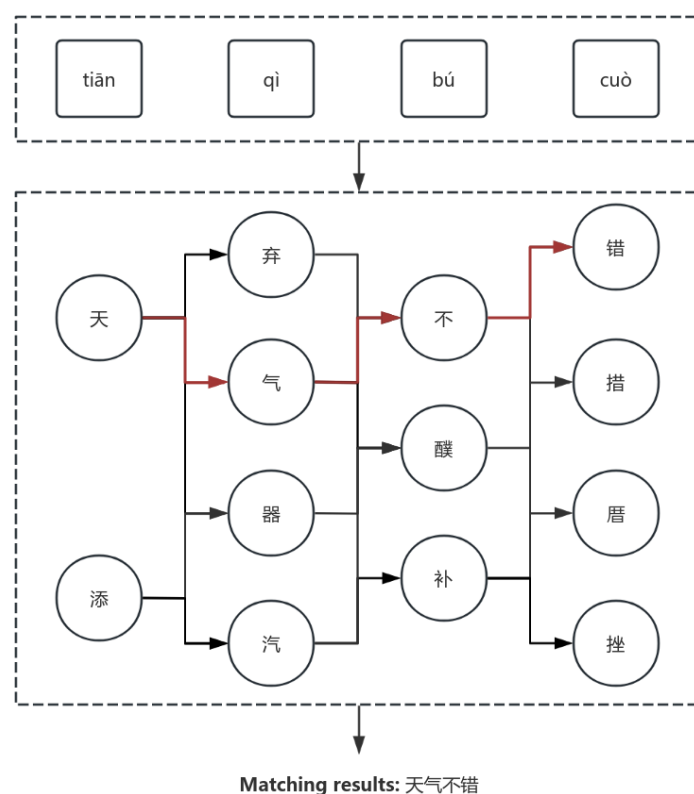
Acoustic models are used in automatic speech recognition systems to represent the relationship between audio signals and phonemes or other language units that make up speech. Modern speech recognition systems use acoustic models and language models to represent the statistical properties of speech. The 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\_etc:** Store the trained acoustic model.

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

**model:** Contains the model\_language folder, the acoustic models AcousticVNet py and LanguageVNet py, and the text required for the acoustic and language models in the model\_language 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] Wang, Y., Yang, T., Liang, H., & Deng, M. (2022). Cell atlas of the immune microenvironment in gastrointestinal cancers: Dendritic cells and beyond. *Frontiers in Immunology*, 13, 1007823.
- [2] Li, X., Wang, J., & Zhang, L. (2025). Gamifying Data Visualization in Smart Cities: Fostering Citizen Engagement in Urban Monitoring. *Authorea Preprints*.
- [3] Wang, J. (2025). Bayesian Optimization for Adaptive Network Reconfiguration in Urban Delivery Systems.
- [4] Li, T. (2025). Enhancing Adverse Event Monitoring and Management in Phase IV Chronic Disease Drug Trials: Applications of Machine Learning.
- [5] Li, X., Wang, J., & Zhang, L. (2025). Named entity recognition for smart city data streams: Enhancing visualization and interaction. *Authorea Preprints*.
- [6] Yang, J. (2025). Application of LightGBM in the Chinese Stock Market.
- [7] Song, X. (2025). User-Centric Internal Tools in E-commerce: Enhancing Operational Efficiency Through AI Integration.
- [8] Wu, W. (2024). Research on cloud infrastructure for large-scale parallel computing in genetic disease.
- [9] Chen, J. (2025). Leveraging Scalable Cloud Infrastructure for Autonomous Driving Data Lakes and Real-Time Decision Making.
- [10] Lin, Y., Liu, J., Cao, Y., Cao, Y., & Wang, Z. (2025). Transfer learning-enhanced modelling of annular aperture arrays and nanohole arrays. *Physica Scripta*, 100(3), 036003.
- [11] Gong, C., Lin, Y., Cao, J., & Wang, J. (2024, October). Research on Enterprise Risk Decision Support System Optimization based on Ensemble Machine Learning. In *Proceeding of the 2024 5th International Conference on Computer Science and Management Technology* (pp. 1003-1007).
- [12] Bohang, L., Li, N., Yang, J. et al. Image steganalysis using active learning and hyperparameter optimization. *Sci Rep* 15, 7340 (2025). <https://doi.org/10.1038/s41598-025-92082-w>