# Review of sPEECH RECOGNITION

12011923 Xudong Zhang

Abstract

Speech recognition refers to the use of computer technology to convert human language into text data that computers can recognize and process. Speech recognition technology is based on speech signal processing, probability statistics, natural language processing and other fields, and is widely used in speech interaction, natural language processing, speech search and other fields. This paper summarizes the background,basic process, the status quo of speech recognition.

**Index Terms—background,basic process, the status quo**

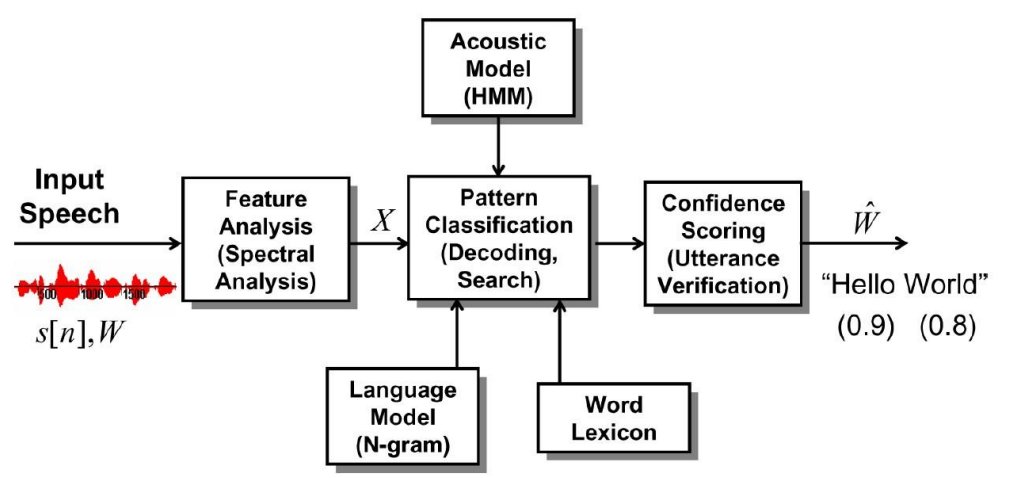
## 1. Introduction

The development of automatic speech recognition technology dates back to the early 1950s. During that time, researchers began trying to build speech recognition systems based on artificial neural networks. However, the development of speech recognition technology was slow due to the limited performance and storage capacity of computers at that time. By the 1970s, speech recognition technology began to mature with the advancement of digital signal processing technology and the improvement of performance of computer. Since then, the emergence and development of a series of artificial intelligence technologies, such as natural language processing, machine learning, and deep learning, have also laid the foundation for the further improvement and wide application of speech recognition technology. In recent years, with the wide application of voice devices such as smart phones and smart speakers, speech recognition technology has entered a new stage of development. At the same time, the fields based on speech recognition technology are also expanding, including smart home, intelligent customer service, unmanned driving, etc. Speech recognition technology is becoming an important bridge connecting people and computers.

Speech recognition technology has become an important branch in the field of artificial intelligence. It not only improves efficiency of production and convenience of life, but also broadens access channels of information and provides basic support for intelligent services, thus driving the development of industrial. Through speech recognition technology, interaction between human and computer becomes more intuitive and natural, and mechanized operation is faster and more efficient. Speech recognition technology has a wide range of application scenarios, including improvement of interaction between human and computer, convenient voice input, processing of natural language, and enhancement of intelligent interaction experience.

At present, there are many commercial products based on speech recognition on the market, covering different fields and application scenarios. Siri is a product developed by Apple, which supports multiple languages and allows users to conduct mobile phone operations, calendar queries, initiating calls and other operations by voice. In addition, there are Amazon Alexa, Xiaodu students, Huawei Xiaobutterfly, iFei input method, Tencent AI Lab and other commercial speech recognition products existing in the market widely. They play their own unique advantages in different application scenarios. With the continuous development and application of speech recognition technology, its influence in the field of economy and society is also increasingly prominent. Its popularity and wide application not only drives the corresponding industrial development and the expansion of the market scale in related fields, but also gives birth to the rise of many innovative enterprises and new forms of business, which plays an important role in promoting the digital transformation and intelligent upgrading of society and provides the direction and support for the digital and intelligent construction of enterprises, cities and governments.

## 2. BASIC PROCESS OF speech RECOGNITION



**Fig. 1**. Overall architecture of a modern ASR system

The basic flow of speech recognition technology includes five main steps: signal preprocessing, feature extraction, acoustic model training, language model training and decoding.

### 2.1 Signal preprocessing

Signal preprocessing is the pre-step of recognition, whose purpose is to process the original speech signal to remove noise, extract speech signal and do other operations, so as to facilitate the subsequent steps. The main steps of signal preprocessing include the following five steps: voice signal acquisition, preprocessing, voice activity detection, framing and feature normalization.

First of all, the voice signal needs to be collected into the computer, which can be collected by microphone, mobile phone and other devices.

Secondly, the speech signal is preprocessed, including denoising, filtering and so on. Noise removing can adopt statistical models and algorithms based on frequency domain or time domain in order to filter out environmental noise, electronic interference, and so on. Filtering can adopt digital filter, wavelet transform and other methods to filter out the noise of low frequency or noise of high frequency.

Thirdly, voice activity detection distinguishes between speech and non-speech parts to determine the start and end points of speech. VAD can adopt algorithms based on features, such as energy, zero crossover rate, and so on. Detection begins when a sound is perceived and ends when the sound disappears or a threshold is reached.

The four step is to divide the speech signal into several frames of equal length. Each frame typically represents 10-30 milliseconds for subsequent steps to extract feature and analysis.

Lastly, the eigenvalues of different samples are normalized to eliminate the scale difference between different samples.

### 2.2 Feature extraction

Feature extraction converts speech signals into mathematical representations for further analysis and processing by machine learning algorithms. The main steps of feature extraction include the following six steps: framing, windowing, Fourier transform, frequency filterbank, DCT and feature superposition.

Framing is to divide the speech signal into short-time frames and analyze the speech signal in each window of time, which can meet the time-varying property of the sound signal.

Windowing is to multiply the frame by the window function of its narrowband filter bank, such as Hanning window, Hamming window, and so on. This can smooth the edge of the frame in the time domain in order to reduce the spectral leakage problem.

After the window is added, the speech signal is time varying. So it is necessary to convert the speech signal in each window into a complex value representation of varying frequency by using the fast Fourier transform.

The frequency domain information of each frame is obtained, and then each frequency point is filtered with a filterbank to enhance the distinction effect of pitch, volume and resonance in speech recognition.

Then, the coefficient of energy of the filter bank is transformed using discrete cosine transform to extract the main frequency components representing the audio signal and remove unnecessary high-frequency noise.

Feature superposition is to combine the features of adjacent frames to capture context information.

### 2.3 Acoustic model training

Acoustic model training is to model and train the speech signal represented by feature vector and its corresponding text annotation data in order to obtain a model that can automatically recognize the speech signal. The main steps of acoustic model training include the following five steps: determination of model structure, initialization of model, model training, evaluation of model and post-processing of model.

Firstly, it is necessary to select the appropriate acoustic model structure and set the hyperparameters of the model according to the dimensions of the feature vector and the distribution of data set.

Then, for some deep learning models, pre-training and initialization of parameters are usually required. Pre-training is to break down the large model into small models to reduce the complexity of the model and make it easier to train.

Model training updates model parameters by means of methods, like maximum likelihood estimation (MLE). So that it can maximize the likelihood probability of the samples of training set. Gradients are usually calculated using the algorithm of backpropagation.

In evaluation of model, validation sets or test sets are often used to evaluate the performance of the model and make adjustments and optimizations. Common evaluation indicators include Accuracy, Recall and F1 value.

Lastly, the acoustic model, language model, decoder and other modules are combined to form a complete speech recognition system. In practical application, make post-processing and optimization, such as error correction of pinyin , keyword weighting and so on, in order to improve the accuracy of recognition.

### 2.4 Language model training

Language model training refers to the use of large-scale text corpus to train a model that can predict the probability of the next word or sentence given previous text input. The purpose of the language model is to evaluate the results of recognition of the acoustic model given a specific audio input, so as to find the optimal results of text transcription and post-process them. The main steps of language model training include the following six steps: corpus data collection and preprocessing, feature extraction, calculation of frequency of word, training language models, evaluation of language model and mergement of model.

Firstly, a large number of language materials are collected for language model training, and then pre-processed by cleaning, de-weighting, word segmentation, and so on.

Feature extraction means that for each language material, the voice activity detection of sentence is used to segment the sentence, and the segmented text is converted into a word or marker of word level.

Then, frequency of word is calculated using the marked corpus, and the frequency of occurrence of each word in the whole corpus is obtained.

In the training language model, given a certain prefix, the standard language model algorithm is used to calculate the probability of a word. The probability of occurrence of all the successors of each prefix is calculated, and the probability is normalized. Then the probability distribution of a word is obtained.

After that, A verification set is used to verify the trained language model, calculate the performance index of the model and adjust the hyperparameters.

Lastly,the trained language model is combined with the acoustic model, decoder and other modules to construct a complete speech recognition system.

### 2.5 Decoding

Decoding is the process of converting a speech signal into a corresponding text sequence. In the stage of decoding, the computer will use acoustic models and language models to calculate the probability of each word or mark, then calculate the score of all possible text sequences and return the most likely text sequence. The main steps of decoding include the following six steps: feature extraction, forward algorithm, scoring of language model , forward-backward algorithm, Viterbi algorithm and post-processing.

Feature extraction refers to the conversion of speech data into feature vectors that can be processed by computers.

Using the feature vector sequence as input to the acoustic model, the acoustic model will generate a probability distribution for each time step. These probability distributions will be used to calculate the score of each speech fragment, which will be used to generate a text sequence of the entire speech.

In the case of accepting a prior sequence of text, the language model calculates the probabilities of the next word or mark, and then uses those probabilities to calculate the score of the final sequence of text.

In order to improve the efficiency of decoding, a forward-backward algorithm is usually used to speed up the decoder. The algorithm uses the method of dynamic programming to calculate the optimal path at each time.

Viterbi algorithm is a decoding algorithm based on dynamic programming. It is a common decoding strategy that can be used to get the final text sequence by passing the probability onto the path with the greatest probability.

In order to improve the accuracy of recognition, post-processing methods, like word graph pruning, can be used to optimize the decoding results.

## 3. The status quo of speech recognition technology

### 3.1 State-of-the-art methods, solutions and algorithms

In the field of speech recognition technology, the most advanced methods, solutions and algorithms mainly include the following five aspects, including deep neural networks, recurrent neural networks, transfer learning techniques, end-to-end speech recognition techniques, use of relevant language models and domain knowledge.

#### 3.1.1 Deep neural networks

Deep neural networks is the most widely used acoustic model, which is a multi-layer neural network structure with the ability to automatically learn complex features. By training a large number of speech training data, deep neural networks can learn more robust and accurate speech representation, and adopt Gaussian mixture model and other methods for modeling to achieve better acoustic model and recognition effect.

#### 3.1.2 Recurrent neural networks

Recurrent neural networks model is a relatively new acoustic model, which is usually used to deal with the problem that the timing features of speech signals are difficult to deal with with traditional methods. Recurrent neural networks can capture the timing and context information of input signals, and introduce memory units into the acoustic model to improve the modeling effect. Variants, like LSTM networks and gated cycle unit networks, can alleviate the problem of the disappearing of gradient and updating slowly during long series training of traditional Recurrent neural network models, and improve the accuracy of speech recognition.

#### 3.1.3 Transfer learning techniques

Transfer learning technology can transfer the model parameters trained on one task to another task or domain to improve the efficiency and accuracy of model training. In speech recognition tasks, transfer learning can improve the effect of model initialization, fine tuning and training by using pre-trained model parameters, accelerate convergence of model and reduce the requirement of data quantity and quality.

#### 3.1.4 End-to-end speech recognition techniques

End-to-end speech recognition technology is an emerging technical framework, which integrates multiple steps into an overall model , such as feature extraction, acoustic model and language model, and uses deep learning neural networks to map speech signals to text sequences directly by bypading the multiple components and complex feature engineering of traditional speech recognition. Because of its simplicity and efficiency, end-to-end speech recognition technology is considered to be the next generation of speech recognition technology.

#### 3.1.5 Use of relevant language models and domain knowledge.

In tasks of speech recognition , the use of more accurate and relevant language models and domain knowledge can greatly improve the accuracy and efficiency of recognition. In order to solve speech recognition problems in different languages and domains, researchers and engineers will use text data sets and domain knowledge in different domains for model training and testing to improve the practicality and versatility of the model application.

#### 3.1.6 Performance comparison.

For performance comparison, factors such as the requirements of different application scenarios and the size and quality of the data set need to be considered. Here are some of the more commonly used performance metrics and how they perform in different ways:

**Accuracy:** Accuracy is the most basic performance indicator, which measures the model's correct recognition rate of input speech. In recent years, with the development of deep learning and end-to-end technology, DNN and end-to-end methods have achieved relatively high accuracy in general speech recognition tasks compared to traditional GMM models and feature-engineering-based methods, especially in larger data sets and more complex application scenarios.

**Efficiency:** Efficiency is an important performance metric that takes into account the cost of time and computational resources required to train and reason a model. Since the end-to-end method bypasses the traditional feature extraction and intermediate steps, additional computation and training time can be reduced, so the end-to-end method has higher efficiency compared to the traditional method and the depth method.

**Adaptability:** The adaptability of the model is very important when facing different domains and tasks. For traditional methods and in-depth methods, specific feature engineering and tuning need to be carried out for different fields and application scenarios, while end-to-end method uses a single model to solve the identification task end-to-end, so it does not need to carry out specific domain knowledge and speech signal analysis and other processing, and has stronger versatility and adaptability.

**Interpretability:** For certain application scenarios, the interpretability of the model is important, such as tasks such as voice diagnostics and security reviews in the medical field. Traditional methods and deep methods can improve the interpretability and visualization of models through specific model structure and feature engineering, while end-to-end methods lack such interpretability and visualization.

### 3.2 Technology challenges and further improvement

The development of speech recognition technology has advanced rapidly, but there are still many challenges. The following are the main challenges of speech recognition technology and the trends in the future:

**Diversity and flexibility:** Speech recognition technology needs to be able to accurately recognize multiple accents, speech speeds, tones, and tones under different ambient noise and audio quality. To solve these problems, machine learning and deep learning algorithms need to be more adaptable to diversity and more flexible, end-to-end speech recognition models.

**Volume of data and quality:** The quality and quantity of voice signal data is very important for training and improving model performance. At present, the voice data set is relatively small, and there is a problem of data imbalance. To add more accurate data, speech recognition technology needs to explore better data enhancement and data synthesis techniques, and there is a need for more and richer speech data sets.

**Multilingual and cross-lingual:** Multilingual speech recognition capabilities will enable users to communicate and use technology in their own language. This will require more research to develop cross-language models robust enough to expand the use of speech recognition technology globally.

**Problem of salience:** Speech recognition technology tends to perform best on the most pronounced, frequent, and high-intensity phonemes, while underperforming on other phonemes of complex nature. This problem can be solved by improving feature extraction and modeling methods.

**Adaptive technology:** Speech recognition technology needs to be able to adapt to the user's habits and special Settings of the device, such as personal accent, audio device quality. This requires more research on online learning and incremental learning to build adaptive, personalized speech recognition models.

In the future, with the development of artificial intelligence and automatic speech recognition technology, speech recognition technology will become more popular and mature, and the application scenarios of speech recognition technology will continue to expand. From mobile phone voice assistants and smart homes, to car voice navigation and phone customer service, voice recognition technology is changing the way people live and work every day. It is expected that future development will pay more attention to the combination and collaborative development of the speech field, such as natural language processing, machine translation, speech synthesis and other technologies, so that speech technology is more diversified and intelligent.

## 4. innovative application, new solution

The development of speech recognition technology has driven the emergence of many innovative applications and new solutions. There are many examples to look at.

**Smart home:** Voice recognition technology can be combined with smart home systems to enable users to control various smart home devices such as lights, temperature, and home entertainment devices by voice.

**Health care:** Voice recognition technology can be applied in the field of health care, such as electronic health records, medical literature transcription, patient interviews, etc., so that medical personnel can enter and manage in the medical process more efficiently, and can also provide rapid response services in emergency situations.

**Online customer service:** Voice recognition technology can be used for online customer service, so that customers can communicate with customer service personnel through voice, improve customer service efficiency and satisfaction.

**Speech translation:** The combination of speech recognition technology and machine translation technology can solve the problem of cross-language communication and make people more convenient to carry out cross-cultural communication.

**Autonomous vehicles:** Voice recognition technology can be combined with autonomous vehicles to enable drivers to operate the vehicle through voice commands, and the vehicle can safely drive through voice recognition technology to recognize road and traffic signals and detect obstacles.

In addition to the above applications, speech recognition technology can also be combined with other technologies, such as emotion and speech recognition, speech recognition and virtual reality, to bring more diversified services and experiences for people.

In order to better meet people's needs, there is also a need for new exploration and improvement in speech recognition technology, some of which include more accurate speech recognition, personalized speech recognition, enhanced speech to text conversion and so on. These innovations will make voice recognition technology smarter and more widely used, which will greatly improve the quality of people's lives and work.

## 5. References

1. Chen, Y., Dai, L., Zhang, H., Gao, Y., & Wu, J. (2021). A Survey on Speech Recognition for Mobile Devices: Advances and Challenges. IEEE Access, 9, 16303-16320. doi: 10.1109/access.2021.3055020
2. Wang, D., Shen, Y., Huang, L., Fang, Y., & Lu, J. (2021). A Review on Artificial Intelligence Based Intelligent Voice Interaction Technology. Proceedings of the 2021 IEEE 11th International Conference on Electronics Information and Emergency Communication (ICEIEC), 53-58. doi: 10.1109/iceiec52512.2021.9461391.
3. Veselý, K., Schwarz, P., Korňák, J., Burget, L., Hannemann, M., Bayer, J., & Černocký, J. (2014). Switchboard recognition using Kaldi: The good, the bad, and the ugly. Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU-14), 563-569.
4. Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., . . . Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6), 82-97.
5. Miller, T., & Tripp, H. (2021). Personalized speech recognition in noisy environments. arXiv preprint arXiv:2103.09377.