Development status, problems and solutions of speech recognition technology

Shuang Zhang
Department of Public Security Information, Liaoning Police College, Dalian, 116036, China
Drew0521@163.com, zhangshuang@lnpc.cn

Abstract. The purpose of automatic speech recognition technology is to enable the machine to "understand" human voice, and convert human voice into text information, which is the key technology to realize human-computer interaction. This paper briefly introduces the development of speech recognition technology and the acoustic model involved, analyzes the problems of noise interference in the environment, the low recognition rate of dialect speech, and the need to improve the speech recognition effect in the far-field environment, and puts forward solutions such as optimizing the acoustic model, building dialect corpus, accurately modeling the sound transmission environment, the future development is prospected.

1. Introduction
Speech recognition takes speech as the research object. It is an important research direction of speech signal processing and a branch of pattern recognition. It involves many fields such as physiology, psychology, linguistics, computer science and signal processing, and even involves human body language (such as the expression and gesture of people when they speak, which can help the other party understand). The ultimate goal of speech recognition is to realize natural language communication between human and machine. In this paper, based on the tracing of the development of speech recognition and the acoustic model involved, the paper analyzes the problems of the current speech recognition technology, and puts forward the optimization and solution of acoustic model.

2. The development of speech recognition technology

2.1 Development history
Speech recognition technology began in the 1950s. Bell Labs successfully developed the first Audry system in the world that can recognize the pronunciation of ten English numerals, which has become the historical beginning of speech recognition technology [1].

Since 1960s, Carnegie Mellon University has carried out continuous speech recognition research, but the development is slow.

In the 1970s, Soviet scientists first proposed to solve the problem of unequal length of speech signals by using dynamic programming method, and on this basis developed a dynamic time warping DTW algorithm [2]. At the same time, the LPC of speech signal linear prediction coding effectively solved the problem of which parameters to extract speech signals as features.

In the 1980s, the statistical model-based method represented by hidden Markov model HMM method gradually occupied a dominant position in speech recognition research [3]. In the late 1980s, ANN, the predecessor of DNN, has become a research direction of speech recognition.
Since the 1990s, the artificial neural network technology has been used as a breakthrough point in speech recognition, which makes speech recognition from theory to practical application. In the 21st century, Hinton [4] proposed the deep confidence network DBN, the research on deep learning has been formally started since then. In the past 10 years, the modeling method based on DNN has become the mainstream voice recognition modeling method. On this basis, speech recognition has been constantly developing and innovating relying on big data technology and cloud computing technology [5].

2.2 Acoustic model of speech recognition
Tracing the development of speech recognition technology, acoustic model is the most influential. The acoustic model involved in each stage is briefly explained.

2.2.1.hidden Markov model (HMM)
Hidden Markov model (HMM) is a statistical model with ordered Markov states, which enables the model to process short-term stationary speech features segmentally and to approximate global non-stationary speech feature sequences. HMM is often used as acoustic model in early text recognition. Because the transition probability of HMM is only related to the previous time, it cannot make full use of context information. There are defects in modeling long-term dependent speech, and the recognition performance will be limited with the increase of data.

2.2.2.Gaussian mixture model (GMM)
GMM is an extension of a single Gaussian probability density function, and GMM can smoothly approximate the density distribution of any shape. In past many years, GMM has been successfully applied in the modeling of speech features and acoustic models of speech recognition [6]. However, it has a serious disadvantage, that is, it cannot effectively model the nonlinear or approximate nonlinear data. For complex speech features, we expect to have better models and better capture speech features.

2.2.3.Artificial neural network (ANN / BP)
Although ANN / BP is accurate in simulating and abstracting human brain functions, they are artificial neural networks, which are just distributed parallel processing models simulating biological perception characteristics. Neural network is widely used in many fields because of its unique advantages, powerful classification ability and input-output mapping ability. But from the current speech recognition system, due to the insufficient description of the time dynamic characteristics of speech signal by ANN / BP, most of them adopt the system combining ANN / BP with traditional recognition algorithm.

2.2.4.Deep neural network hidden Markov (DNN-HMM)
Many of the traditional machine learning models belong to shallow structure, such as GMM, HMM, ANN, BP, etc. they can not effectively learn the complex structural information in the signal, and have certain limitations on the expression ability of complex signals. The deep structure model is more suitable for dealing with complex types of signals, because the deep structure has multi-layer nonlinear transformation [7], and has stronger expression and modeling ability.

Since 2011, the DNN-HMM acoustic model has achieved a larger and consistent effect than the traditional GMM-HMM acoustic model in multi language and multi task speech recognition. Compared with the traditional speech recognition system based on GMM-HMM, the biggest change is to replace GMM model with deep neural network to model the observation probability of speech. Compared with GMM, DNN has the following advantages: 1. Using DNN to estimate the posterior probability distribution of HMM States does not need to assume the distribution of voice data; 2. The input features of DNN can be the fusion of multiple features, including discrete or continuous; 3. DNN can use the structural information contained in adjacent voice frames. However, the disadvantage lies in the lack of flexibility in modeling the longer time correlation of time series information. The model of DNN-HMM recognition system is shown in Fig. 1.
2.2.5. Recurrent neural network (RNN)

Considering the long-term correlation of speech signal, we will think of choosing the neural network model with stronger long-term modeling ability. Therefore, in recent years, recurrent neural network has gradually replaced the traditional DNN to become the mainstream voice recognition modeling scheme. The network structure of DNN and RNN is shown in Fig. 2.

RNN adds a feedback connection on the hidden layer, which is also the biggest difference between RNN and DNN. This means that part of the input of the hidden layer at the current moment is the output of the hidden layer at the previous time, which enables the RNN to see the information of all the previous moments through the circular feedback connection, which endows the RNN with the memory function. These characteristics make RNN very suitable for modeling time series signals.

However, the traditional RNN has the problem of gradient disappearing in the training process, which makes the model difficult to train. In order to overcome the problem of loss of gradient, some researchers proposed long-term and long-term memory LSTM-RNN [8]. On this basis, a bi-directional LSTM-RNN (BLSTM-RNN) is proposed. When processing the current frame, it can use the historical and future
voice information to make more accurate decisions and achieve better performance improvement than the unidirectional LSTM.

2.2.6 Convolutional neural network (CNN)
The core of CNN is convolution operation (or convolution layer), which is another model that can effectively utilize long-term context information [9]. CNN has been used in speech recognition system as early as 2012, and many researchers have been actively engaged in the research of CNN based speech recognition system, but there has been no big breakthrough. The main reason is that they have not broken through the traditional feedforward neural network which uses fixed length frame splicing as input, so they can not see enough speech context information. Another drawback is that they only regard CNN as a feature extractor, so they use only one or two convolution layers, which is very limited. Aiming at these problems, iFLYTEK proposed a speech recognition framework based on deep full sequence convolutional neural network (DFCNN), which uses a large number of convolution layers to directly model the whole sentence speech signal, so as to better express the long-term correlation of speech.

3. Problems faced by speech recognition technology

3.1 Noise interference in the environment
When speech recognition system is used in noisy environment, the speaker will have emotional or psychological changes, which will lead to distortion of pronunciation, change of pronunciation speed and tone. In this case, the accurate recognition of speech is a big problem. At present, the published accuracy rate of speech recognition is 97%, which can only be achieved when the indoor environment is relatively quiet. But in fact, such a quiet environment is often not easy to achieve.

3.2 The recognition rate of dialect speech is not high
Speech recognition has a high recognition accuracy for Mandarin, but it may be difficult for speech recognition with dialects. Due to the differences in age, gender, accent, dialect, speed, intensity and habit of pronunciation, it is still not easy to realize the system to automatically adapt to the characteristics of most people's voice lines and eliminate these differences to achieve stable speech recognition.

3.3 Speech recognition in far-field environment needs to be improved
The challenge of far-field identification is mainly caused by the complex signal propagation environment. In the case of far-field identification, the location of sound source is uncertain, there are many noise sources, the noise is large, and the signal-to-noise ratio decreases sharply. In recent years, with the development of far-field pickup technology, microphone array layout and software algorithm are more and more rich, and the far-field pickup ability is improved significantly. But even so, it still faces great challenges, especially in the background noise environment, the effect still has a lot of room for improvement.

4. Solutions

4.1 Optimization of acoustic model
Based on the existing acoustic model of deep full sequence speech recognition, the model is optimized. It uses the invariance of convolution to overcome the diversity of speech signal itself, and adjusts the multi-layer convolution unit structure of the model. The unit structure is composed of convolution layer, activation function, dropout and pooling layer, that is, dropout is added between convolution layer, pooling layer and convolution layer When the amount of training data is large, the number of neurons in hidden layer is modified to prevent the over fitting of the network, which enhances the generalization ability and robustness of the neural network, and then enhances the robustness of the speech recognition system.
4.2 Construction of dialect corpus
Taking local dialects from all over the country as the object, according to a series of measures, such as the choice of speakers (gender, age, place of birth, place of residence), corpus design, recording standards, data storage standards, corpus tagging standards, corpus evaluation norms and other measures, dialect phonetic samples are collected and dialect database is created. Using dialect corpus to train the acoustic model, improve the accuracy of dialect speech recognition and enhance the ability of dialect speech recognition.

4.3 Accurate modeling of sound propagation environment
For far-field speech recognition, the most basic work we need to do is to model the propagation environment more accurately. This model is not only helpful for us to understand the attenuation characteristics of the signal, but also to design a targeted speech signal enhancement algorithm. Moreover, a large number of far-field speech data can be generated quickly by using the model for acoustic model training at the recognition end, which helps to solve the problem of difficult acquisition of far-field voice data.

5. Future development direction
In order to solve the problem of speech recognition and make the performance of speech recognition system close to human level in all cases, new acoustic modeling technology is needed. The next generation speech recognition system should have the following features:

Firstly, a dynamic system with many interconnected components and loop feedback can always predict, modify and adapt;

Secondly, they can do better in semantic understanding;

Finally, we can learn the key pronunciation features from the training set and generalize them to unknown speakers, accented voice and noisy environment [10].

6. Summary
This paper briefly reviews the history of the development of speech recognition technology, introduces the acoustic model of speech recognition, discusses the bottleneck problems in the current research, puts forward solutions, and prospects the future development, in order to draw on the current research, cause everyone's thinking, and make a breakthrough in technology, solve the bottleneck, and make greater progress in speech recognition.

Acknowledgement
Thanks: Fund Project: Key research and development (R&D) project of Liaoning Provincial Science and Technology Department, No 2020JH2/10300109

Reference
[1] DAVIS K. H, BIDDULPH R, BALASHEK S. Automatic recognition of spoken digits[J]. Journal of the Acoustical Society of America, 1952, 24(6): 637
[2] FERGUSON J D. Application of hidden Markov models to text and speech[EB]. 1980.
[3] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Readings in Speech Recognition, 1990, 77(2): 267-296.
[4] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7):1527-1554.
[5] YU D, DENG L. Deep learning and its applications to signal and information processing[J]. IEEE Signal Processing Magazine, 2011, 28(1): 145-154.
[6] (USA) Yu Dong, Deng Li, Yu Kai, Qian Yanmin. Analysis of deep learning speech recognition practice [M]. Beijing: Electronic Industry Press, 2019:15
[7] HINTON G E. Training products of experts by minimizing contrastive divergence[J]. Neural Computation, 2002, 14(8):1771-1800.
[8] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[9] LECUN Y, BENGIO Y. Convolutional networks for images, speech and time-series[M]. Cambridge: MIT Press, 1995.

[10] (USA) Yu Dong, Deng Li, Yu Kai, Qian Yanmin. Analysis of deep learning speech recognition practice [M]. Beijing: Electronic Industry Press, 2019: 254