Review of Automatic Speech Recognition (June 2019)

Jiahao Hu. Author

Abstract—**This paper first introduces the dvelopment and current situation of automatic speech recognition, then introduces classification of speech recognition system. Also, it introduces several methods of speech recognition and speech recognition based on deep learning in detail. Moreover, the structure of speech recognition system is mentioned. Finally, this paper describes the future research direction of speech recognition technology.**

I. INTRODUCTION

Speech is a kind of voice which represents a certain meaning and is the material shell of language. Speech is the most natural, basic and most important information carrier used by human beings when they communicate with each other. In today's highly informationized society, a series of voice processing technologies and their applications have become an indispensable part of the information society.

Pronunciation is a complex process, including a series of psychological and physiological actions. When people need to express certain information through voice, first of all, this information is expressed in the speaker's brain in some abstract form, and then converted into a group of nerve signals, which act on the vocal organs to produce voice signals carrying information.

Speech recognition technology is a technology that enables machines to transform human speech signals into corresponding text or commands through the process of recognition and understanding[1]. It belongs to the category of multi-dimensional pattern recognition and intelligent computer interface. It is a comprehensive technology in the fields of acoustics, semantics, computer, information processing, artificial intelligence and so on. It has been widely used in industry, military, transportation, civil and other fields. The goal of speech recognition is to make computers "understand" the language spoken by human beings. With the development of speech recognition technology, various speech products have emerged. Speech recognition products have occupied an increasing proportion in human-computer interaction applications.

II. DEVELOPMENT AND CURRENT SITUATION

The research of speech recognition can be traced back to the 1950s. In 1952, Bell Laboratory's Audrey system was the first speech recognition system to recognize ten English digits.

In the late 1960s and early 1970s, several basic ideas of speech recognition emerged. Among them, LPC and DTW technologies were proposed, which effectively solved the problems of feature extraction and unequal length speech matching. Vector Quantization (VQ) and Hidden Markov Model (HMM) theory have been applied in practice, and a speaker-specific isolated speech recognition system based on linear prediction Cepstrum and DTW technology has been preliminarily realized.

In the 1980s, with the successful application of HMM model and artificial neural network (ANN) in speech recognition, people finally broke through the three major obstacles of speech recognition in the laboratory: large vocabulary, continuous speech and non-specific person. At the level of acoustic recognition, based on large-scale speech data of multiple speakers, the phoneme recognition rate has made great progress by modeling HMM of context pronunciation variants in continuous speech. At the linguistic level, on the basis of large-scale corpus, the fuzziness of homonyms and near-syllables caused by recognition can be effectively distinguished by counting the correlation between two or three adjacent words. In addition, combined with efficient and fast search algorithm, real-time continuous speech recognition system can be realized.

In the 1990s, people began to further study the combination of speech recognition and natural language processing, and gradually developed into a man-machine dialogue system based on natural spoken language recognition and understanding. Artificial Neural Network (ANN) and HMM model are combined to improve the recognition rate and robustness of the system.

In 2006, due to the successful application of deep learning theory in machine learning, people began to pay attention to it. In the next few years, the research hotspot in the field of machine learning began to gradually turn to in-depth learning. Deep learning model generally refers to a deeper structural model, which has more layers of non-linear transformation than the traditional shallow model, is more powerful in expression and modeling ability, and has more advantages in complex signal processing. The advantages of in-depth learning have attracted the attention of many researchers in the field of speech signal processing, and people have begun to study it actively. In the next few years, through the unremitting efforts of researchers, many breakthroughs have been made. In 2009, in-depth learning was applied to speech recognition for the first time. Compared with hidden Markov model speech recognition system, it achieved more than 20% relative performance improvement. Since then, the acoustic model based on deep neural network has gradually replaced GMM as the mainstream model of speech recognition acoustic modeling, and greatly promoted the development of speech recognition technology, breaking through the bottleneck of speech recognition performance requirements in some practical application scenarios, and making speech recognition technology truly practical[2].

III. Classification of Speech Recognition System

Speech recognition has different classification methods from different perspectives and requirements.

According to the size of vocabulary, it can be divided into small vocabulary speech recognition system, medium vocabulary speech recognition system and large vocabulary speech recognition system. Generally speaking, with the increase of vocabulary in the vocabulary, the confusion between words increases, and the recognition rate of the system decreases.

According to the limited range of speaker, it can be divided into specific person recognition system and non-specific person voice system. For speech recognition of a specific person, a large number of pronunciation data must be input by the user and trained before use. In speaker-independent recognition, the user does not need to input a large number of training data in advance, and the speech signal variability is very large. Therefore, speaker-independent speech recognition system should learn the basic features of Speaker-Independent pronunciation speed, speech intensity, and pronunciation mode from a large number of different speaker's pronunciation samples, and find and summarize their similarities as recognition criteria. This learning and training process is quite complex. The speech samples used should be collected in advance and completed before the system is generated.

According to the way of pronunciation, it can be divided into isolated word recognition system, conjunction recognition system, continuous speech recognition system and speech understanding system. The first three systems recognize single vocabulary, multiple consecutive vocabulary and normal spoken sentences. Language understanding system is based on speech recognition, using linguistic knowledge to infer the meaning of speech.

IV. Several Methods of Speech Recognition

A. Dynamic Time Warping (DTW)

Endpoint detection of speech signal is a basic step in speech recognition. It is the basis of feature training and recognition. The so-called endpoint detection is to exclude the silent segment from the voice signal by locating the beginning and end points of various segments in the voice signal. In the early stage, the main basis of endpoint detection is energy, amplitude and zero-crossing rate. But the effect is often not obvious. In the 1960s, Itakura, a Japanese scholar, proposed a dynamic time warping algorithm. The idea of the algorithm is to extend or shorten the unknowns evenly until they are consistent with the length of the reference mode. In this process, the time axis of unknown words is distorted or bent unevenly, so that their characteristics are aligned with those of the model. Continuous speech recognition is still the mainstream method. At the same time, many improved DTW algorithms have been proposed in the isolated word recognition system.

B. Hidden Markov Model (HMM)

Hidden Markov Model (HMM) was introduced into speech recognition theory in 1970s. Its appearance has made a substantial breakthrough in natural speech recognition system. At present, most speaker-independent speech recognition systems with large vocabulary and continuous speech are based on HMM model. HMM establishes a statistical model for the time series structure of speech signals and regards it as a mathematical double stochastic process: one is to simulate the implicit stochastic process of the change of statistical characteristics of speech signals using Markov chains with finite number of states, the other is the stochastic process of the observation sequence associated with each state of Markov chains. The former is shown by the latter, but the specific parameters of the former are unmeasurable. Speech process is actually a double random process. Speech signal itself is an observable time-varying sequence, which is a parameter flow of phonemes produced by the brain according to grammatical knowledge and speech needs (unobservable state)[3]. It can be seen that HMM reasonably imitates this process and describes the overall non-stationarity and local stationarity of speech signals well. It is an ideal speech model.

C. Vector Quantization (VQ)

Vector quantization is an important method of signal compression. Compared with HMM, VQ is mainly suitable for speech recognition with small vocabulary and isolated words. The process is to make every frame of speech signal waveform into a vector in multi-dimensional space, and then quantize the vector. In quantization, the infinite multidimensional space is divided into several regions, and then the input vectors are compared with these boundaries, and quantized as the central vector value of the boundary of the region with the smallest distance. The design of vector quantizer is to train a good codebook from a large number of signal samples, to find a good definition formula of distortion measure from the actual effect, and to achieve the maximum possible average signal-to-noise ratio with the least amount of search and calculation of distortion.

D. Support Vector Machine (SVM)

*S*upport vector machine (SVM) is a learning machine model based on statistical theory. It overcomes the shortcomings of traditional empirical risk minimization method by using structural risk minimization principle. Considering both training errors and generalization ability, it has many advantages in solving small samples, non-linearity and high-dimensional pattern recognition, and has been applied to the field of pattern recognition.

E. Artificial Neural Network (ANN)

Artificial neural network (ANN) has been a research hotspot in the field of artificial intelligence since 1980s. It abstracts the human brain neuron network from the perspective of information processing, establishes a simple model, and forms different networks according to different connection modes. In engineering and academia, it is often referred to as neural network or similar neural network. Neural network is an operation model, which consists of a large number of nodes (or neurons) connected with each other. Each node represents a specific output function, called an excitation function. The connection between two nodes represents a weighted value for the signal passing through the connection, which is called the weight, which is equivalent to the memory of the artificial neural network. The output of the network varies according to the connection mode, weight value and excitation function of the network. The network itself is usually an approximation of some algorithm or function in nature, or it may be an expression of a logical strategy. In recent ten years, the research work of artificial neural network has been deepening continuously, and great progress has been made. In the fields of pattern recognition, intelligent robots, automatic control, prediction and estimation, biology, medicine, economy and so on, it has successfully solved many practical problems which are difficult to be solved by modern computers, showing good intelligence characteristics.

In addition to the several speech recognition methods mentioned above, the most popular one in recent years is the acoustic model based on deep learning.

V. Speech Recognition Technology Based on Deep Learning

Deep learning is one of the technology and research fields of machine learning. Artificial intelligence is realized in computer system by establishing artificial neural network with hierarchical structure. Because hierarchical ANN can extract and filter input information layer by layer, in-depth learning has the ability to represent learning, and can achieve end-to-end supervised learning and unsupervised learning. In addition, in-depth learning can also participate in the construction of reinforcement learning system, forming in-depth reinforcement learning.

The class ANN used in deep learning has many forms, and its complexity is commonly called "depth". Depending on the type of architecture, the form of deep learning includes multi-layer perceptron, convolutional neural network, cyclic neural network, depth confidence network and other hybrid architecture. In-depth learning uses data to update the parameters in its construction to achieve training objectives. This process is commonly referred to as "learning". The common learning methods are gradient descent algorithm and its variants. Some statistical learning theories are used to optimize the learning process.

A deep neural network (DNN) is an artificial neural network (ANN) with multiple layers between the input and output layers. The DNN finds the correct mathematical manipulation to turn the input into the output, whether it be a linear relationship or a non-linear relationship. The network moves through the layers calculating the probability of each output. For example, a DNN that is trained to recognize dog breeds will go over the given image and calculate the probability that the dog in the image is a certain breed. The user can review the results and select which probabilities the network should display and return the proposed label. Each mathematical manipulation as such is considered a layer, and complex DNN have many layers, hence the name "deep" networks.

DNNs can model complex non-linear relationships. DNN architectures generate compositional models where the object is expressed as a layered composition of primitives. The extra layers enable composition of features from lower layers, potentially modeling complex data with fewer units than a similarly performing shallow network.

Compared with the traditional GMM-HMM-based speech recognition framework, the biggest change in the DNN-HMM-based speech recognition acoustic model framework which is shown in Figure 1 is to use DNN instead of GMM to model the observation probability of speech. The advantages of DNN over GMM are as follows: 1. Estimating the posterior probability distribution of HMM state using DNN does not require assuming the distribution of voice data. 2. The input features of DNN can be a fusion of multiple features, including discrete or continuous 3. DNN can utilize the structural information contained in adjacent voice frames[4].

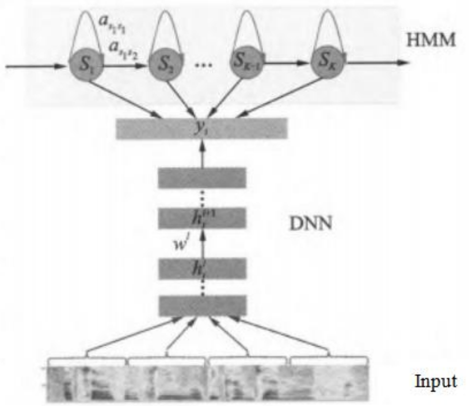

Fig. 1 The DNN-HMM Speech Recognition Block

VI. Structure of Speech Recognition System

Speech recognition is essentially a process of pattern recognition. Its basic principle block diagram is shown in Figure 2. It mainly includes several functional modules such as speech signal preprocessing, feature extraction, feature modeling, similarity measurement and postprocessing, among which the post-processing module is optional.

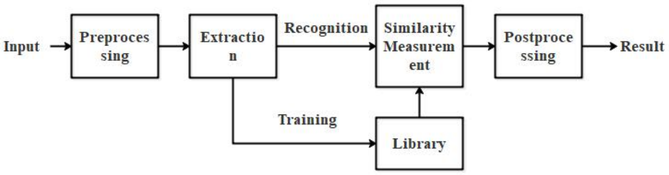

Fig. 2. The Speech Recognition System

In the preprocessing module, the input original speech signal is processed, the unimportant information and background noise are filtered out, and the endpoint detection, speech sub-frame and pre-emphasis of the speech signal are processed.

The feature extraction module is responsible for calculating the acoustic parameters of speech and calculating the features, so as to extract the key feature parameters reflecting the signal characteristics, so as to reduce the dimension and facilitate subsequent processing.

In the training phase, the user input several training voices. After preprocessing and feature extraction, the system obtains the feature vector parameters and establishes or modifies the reference mode library of the training voice.

In the recognition stage, the feature vector parameters of the input speech and the patterns in the reference pattern library are compared to measure the similarity, and the category of the pattern with the highest similarity is used as the intermediate candidate for recognition.

The post-processing module continues to process the candidate recognition results, and gets the final recognition results through the constraints of language model, lexical, syntactic and semantic information.

VII. FURTHER RESEARCH AND CONCLUSION

At present, speech recognition technology based on deep learning has made great progress compared with traditional GMM-HMM technology. However, there is still much room for research on anti-noise, far-field recognition, multilingual mixing and emotional recognition.

Speech recognition technology in quiet environment has reached a practical level. However, in some special environments, such as strong noise interference or far-field situation, the performance of speech recognition system still does not meet the practical requirements. To solve the problem of speech recognition under far-field and strong noise interference is a problem that needs further study.

In the field of multilingual hybrid recognition and infinite vocabulary recognition, future speech and acoustic models may incorporate multilingualism, so users do not have to switch back and forth between languages. In addition, the further improvement of the acoustic model and the semantics-based language model can also help users to recognize infinite vocabulary as little as possible or not affected by vocabulary. With the perfect combination of speech recognition technology, machine translation technology and speech synthesis technology, people who speak different languages all over the world can communicate freely in real time without language barriers.

In addition, the combination of emotional intelligence and computer technology has produced the research topic of emotional computing, which will greatly promote the development of computer technology. Automatic emotion recognition is the first step to emotional computing. As the most important medium of human communication, voice carries abundant emotional information. How to automatically recognize the speaker's emotional state from speech is also widely concerned by researchers in various fields.

REFERENCES

[1] H. edited by Marcus, *Advanced speech recognition : concepts and case studies*. Jersey City, NJ: Jersey City, NJ : Clanrye International, 2015, 2015.

[2] D. Yu, *Automatic Speech Recognition [electronic resource] : A Deep Learning Approach*. London : Springer London : Imprint: Springer, 2015, 2015.

[3] M. S. V. edited by David R. Westhead, *Hidden Markov Models [electronic resource] : Methods and Protocols*. New York, NY : Springer New York : Imprint: Humana Press, 2017, 2017.

[4] J. Novoa, J. Fredes, V. Poblete, and N. B. Yoma, "Uncertainty weighting and propagation in DNN–HMM-based speech recognition," *Computer Speech & Language,* vol. 47, pp. 30-46, 2018.