Deep Learning in Automatic Speech Recognition (ASR): A Review

Taiyao Zeng(✉)

Bachelor of Science in Information Technology,
University of Technology Sydney, Sydney 2007 511727, Australia
13397183@student.uts.edu.au

Abstract. In today's big data era, many traditional machine learning algorithms for processing large amount of raw unlabeled speech data are no longer applicable. At the same time, deep learning models, with their powerful modeling capability for massive data, are able to process unlabeled data directly and have become a hot research topic in the field of automatic speech recognition. The paper gives an overview of the application of deep learning in speech recognition. The research results of deep learning in the field of speech recognition in recent years are introduced, the correlation between traditional speech recognition models and the current deep learning models is discussed, the development trend of deep learning in the field of speech recognition is analyzed, and it is pointed out that deep learning models need to absorb the ideas of traditional speech recognition models in order to better build a speech recognition system based on deep learning models.

Keywords: Deep Learning · Automatic Speech Recognition · CNN · LAS

1 Introduction

Before the birth of human writing, language has existed for a long time as a tool for people's communication and communication. It is an important bridge and medium for the communication of human civilizations and promotes the development of civilization. Until now, in the context of the rapid development of information technologies such as big data, the Internet of Things, and cloud computing, artificial intelligence technology based on deep learning has also developed rapidly and has achieved a huge transformation from the theoretical research level to the practical application level. With the support of artificial intelligence technology, the automatic speech recognition system has also developed from "unusable" to "available", showing very high application value and good development prospects [1]. The traditional automatic speech recognition model has a complex structure and requires a lot of computing and storage resources, making it difficult for automatic speech recognition systems to enter the life of ordinary people. However, with the introduction of deep learning into the field of speech recognition, the complexity of training models is greatly reduced, and even A deep learning automatic speech recognition model can be loaded into a mobile device, such as Google's voice assistant model with a size of 80M [2]. It can be seen that the rapid development of the

C. F. Peng et al. (Eds.): MMET 2022, ASSEHR 703, pp. 173–179, 2023.
https://doi.org/10.2991/978-2-494069-51-0_23

field of automatic speech recognition is inseparable from the application of deep learning models in the field of language recognition. The research goal of this paper is the application and development trend of deep learning models in the field of speech recognition, and through research to reveal deep learning and traditional automatic speech recognition technology relationship, explore the main research directions and methods of deep learning technology in the field of speech recognition, and compare the advantages and limitations of each direction.

2 Deep Learning

Deep learning is a branch in the field of machine learning research. It can be understood as the development of artificial neural networks. It is essentially a method of training deep structural models, and it is also an algorithm for modeling complex relationships between data through multiple layers [3], which is currently used in speech and image recognition, machine translation, and social filtering. In terms of the structure of the model, it belongs to the deep structure, most of the current learning algorithms such as regression and classification belong to simple learning, GMM model, HMM model, support vector machine and multi-layer perceptron belong to shallow structure, support vector machine this classification model is the most widely used, the common feature of the shallow structure is: through the simple structure can convert the input signal, features to the feature space of a particular problem The common feature of the shallow structure is that it can transform the input signal and features into the feature space of a specific problem by the simple structure, but it is difficult to express complex functions, it has some limitations in the processing of natural signals (human speech, natural images) [4].

3 Automatic Speech Recognition

Automatic speech recognition simply means that the machine understands human speech and realizes that the human voice gives commands directly to the computer, and the computer executes the commands according to the recognized and processed voice, thus realizing the intelligent interaction between human and computer [5]. Most of the traditional automatic speech recognition models use Hidden Markov Gaussian Mixture Model (HMM-GMM) [6], which is based on the theory of likelihood and probability. Figure 1 shows the structure of the traditional speech recognition model, which is more complex and consists of several different components. The automatic traditional automatic speech recognition model needs to use the relationship between the speech signal and the digital model in the front-end speech preprocessing stage and needs to use the combination of sampled speech signals to predict the signal, which needs to use the linear prediction analysis method [2]. However, the complexity of speech information extraction due to the different pronunciations of people with different languages, genders and ages makes it difficult to adapt the preprocessing models in traditional speech recognition to these different scenarios.

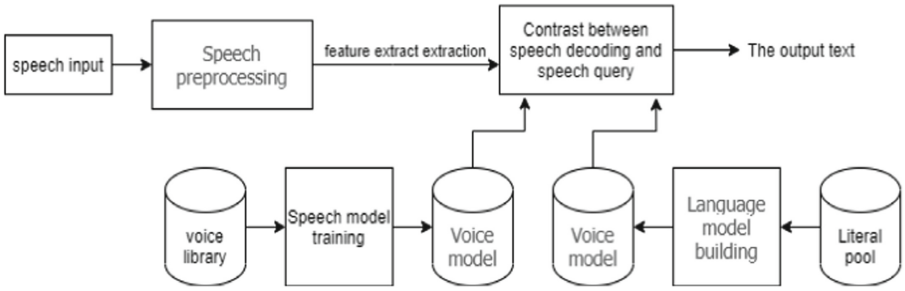

Fig. 1. Traditional automatic speech recognition

Deep learning can make full use of the correlation between features and merge speech features from consecutive frames for training, allowing speech recognition systems to increase recognition rates substantially [7]. In the early stages of combining deep learning with automatic speech recognition, researchers added deep learning models to the structure of Fig. 1, using deep learning models to optimize data for automatic speech recognition, such as Tandem's approach used to add deep learning models after front-end speech preprocessing, making the output feature values more favorable for the speech decoding stage. This approach does not destroy the original automatic speech recognition model. Next, researchers experimented with incorporating deep learning models into existing components, such as the DNN-HMM Hybrid approach that uses deep learning models to replace the GMM structure in the HMM, combining deep learning models with the language decoding component. Then later, researchers tried to replace components of traditional automatic speech recognition models with deep learning models, such as using deep learning models to implement speech models in traditional automatic speech recognition, and using deep learning models to implement language models. Finally, deep learning models were used to replace the entire automatic speech recognition model.

4 Deep Learning Models in Automatic Speech Recognition

In this section, it will introduce deep learning in speech recognition, a landmark of some model structures and discuss their advantages and disadvantages.

CNN (Convolutional neural networks) is the first algorithm for learning a true multilayer neural network structure, building a deep convolutional neural network with input, convolutional, pooling, fully connected, and output layers [3, 9]. Since training deep neural networks with BP algorithms was almost impossible before the emergence of unsupervised layer-by-layer greedy pre-training of deep neural networks, convolutional neural networks are a special case, which is used to reduce the number of spatial relationship parameters to improve the general forward BP training speed [3].

Convolutional neural networks are introduced to continuous speech recognition and compared with the commonly used deep neural networks. Convolutional neural networks have a better physical meaning compared to deep neural networks by observing local features through convolutional layers and then integrating the information in full network

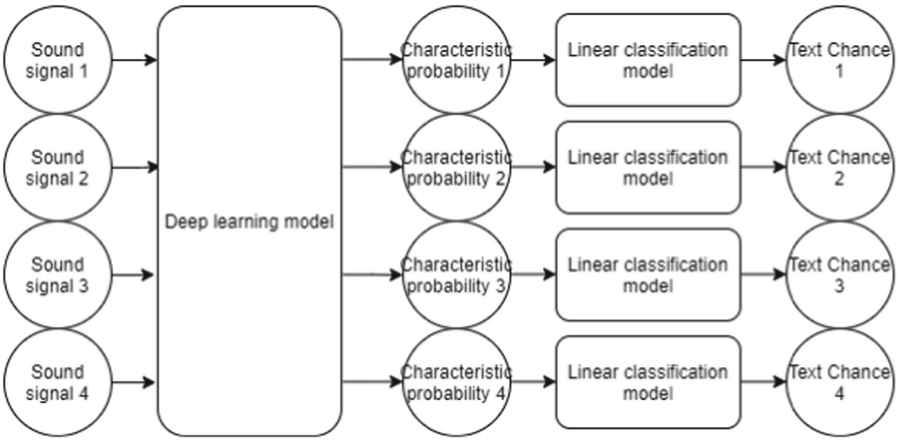

Fig. 2. CTC Model

layers to finally obtain the output probability [8]. Also, due to the weight sharing of convolutional neural networks, the model complexity is greatly reduced.

Connectionist Temporal Classification (CTC) is a neural network-based temporal class classification model. CTC can achieve online learning, i.e., when the speech input is received, it can get the corresponding output without waiting for the entire speech segment to be input. In traditional speech recognition systems, acoustic model training is supervised learning, and it is necessary to know the labeled output corresponding to the input of each speech frame in order to conduct effective training, and the training data preparation stage must align the speech with the label, and it is a difficult task to correspond each input and output of the speech [2]. However, CTC introduces blank symbols to solve the problem of unequal input and output sequences, the main idea is to maximize the sum of all possible corresponding sequence probabilities, without considering the alignment of speech frames and characters, only the input and output can be trained [10]. For a sequence data, it is easier to determine the corresponding labeled pronunciation. As Fig. 2 shows the structure of the CTC model, the deep learning model uses an RNN and the output of the text is performed by a linear classification model.

CTC has two advantages: firstly, it does not need to align data and annotations one by one alignment; the second is that CTC directly outputs the probability of sequence prediction without additional processing.

The third model is the RNN Transducer (RNN-T) structure, which provides language model modeling capability, which can jointly optimize speech models with language models to facilitate online speech recognition.

The deep learning model of the RNN-T model has two components, an RNN structure implementing the structure of the acoustic model and another RNN implementing the structure of the language model [2]. The RNN-Transducer model is actually an improvement on the CTC model, because the CTC model for acoustic modeling has two serious bottlenecks, one is the lack of language modeling capability to integrate language models for joint optimization, and the other is the inability to model the dependencies between model outputs [11]. RNN-Transducer addresses the shortcomings of CTC,

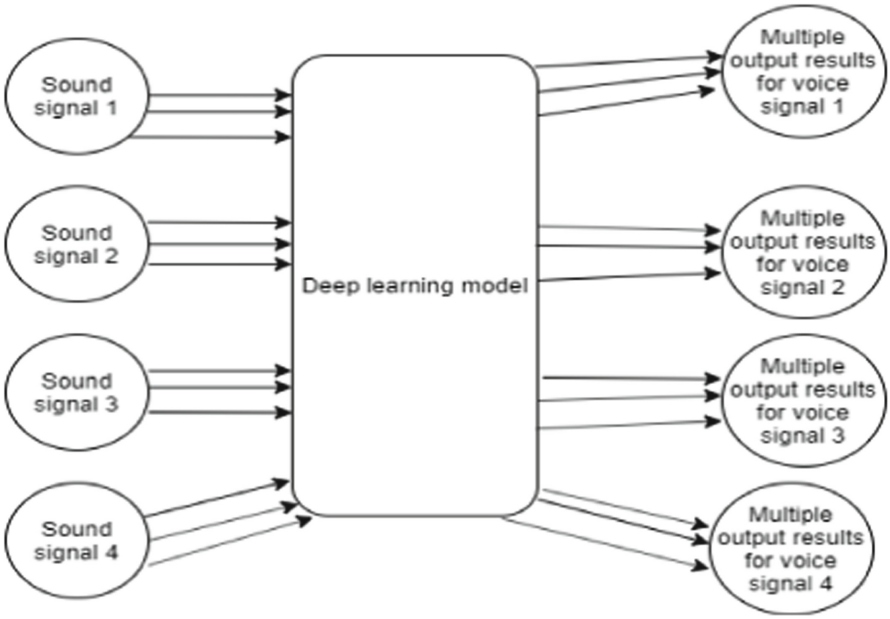

Fig. 3. Input and output of RNN-T Model

making the model more suitable for speech tasks with outstanding advantages such as end-to-end joint optimization, language modeling capability, and ease of implementing Online automatic speech recognition.

As shown in Fig. 3, the RNN-T is used in the input data through the acoustic model output is not one-to-one correspondence, but one input result can have The number of output results is controlled by the language model.

The fourth model is Listen, Attend, and Spell (LAS), which consists of two main parts: 1. Listener, which is Encoder, uses multi layer RNN to extract hidden features from the input sequence; 2. Attend and Spell which means that Attention is used to get the context vector, The LAS model is truly end-to-end, with all components trained jointly, and no independence assumptions required. However, the LAS model needs to recognize the whole input sequence afterwards, so the real-time performance is poor, and the model has been improved by many scholars since then [10]. Its structure block diagram is shown in Fig. 4.

It is a model structure that converts speech into characters completely using deep learning model. LAS does not apply deep learning on top of the traditional DNN-HMM model, but completely uses deep learning models to implement speech recognition.

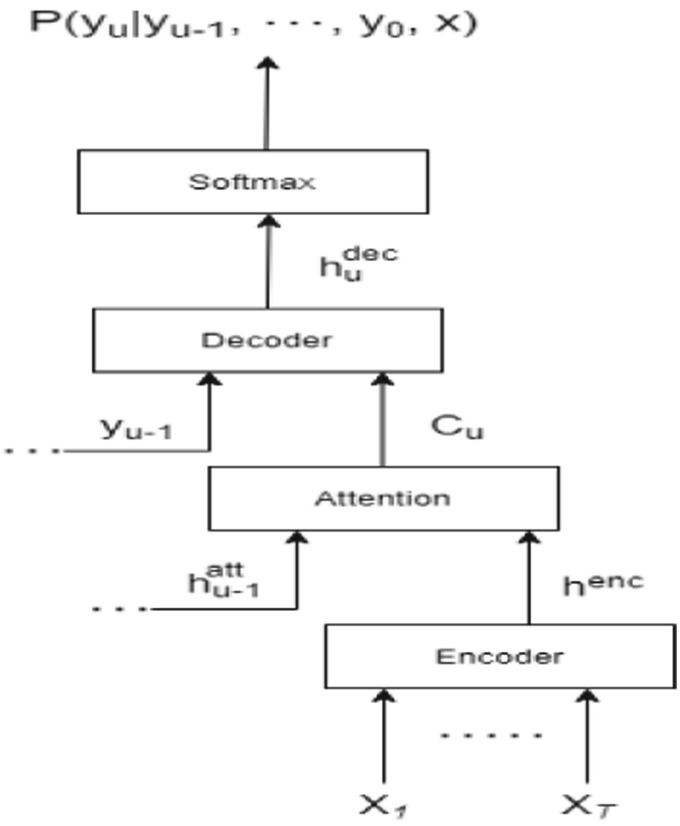

Fig. 4. LAS model framework diagram

5 Conculsions

The introduction of deep learning into the field of automatic speech recognition started with optimizing the output of one part of the HMM model, such as Tandem, then deep learning models were used to complete the function of one structure of the HMM model, such as DNN-HMM Hybrid, then evolved to enable deep learning models to complete the core function of the HMM structure, such as CTC, and recently evolved to use deep learning models to implement the whole automatic speech recognition structure, such as LAS. LAS is the main direction of academic research nowadays, but in the process, the relevant results of traditional HMM models are also applied to assist the training of LAS models to improve the accuracy of speech recognition of the models. There are two research approaches to use deep learning models in the field of speech recognition. The first direction is to apply new deep learning models to existing speech models, and the other research direction is to build a speech recognition model composed entirely of deep models.

References

1. Yawen, G. (2022) Research on Speech Recognition Methods in the Context of Deep Learning for Artificial Intelligence. Software, 43(05):122-124.
2. Jia, W. Dongmei, L. (2020) A review of deep learning applications in speech recognition. Computer Knowledge and Technology, 13(16): 191-197.
3. Yimin, H. Huiqiong, Z. Zhengyi, W. (2017) A Review of Research Advances in Deep Learning for Speech Recognition. Application Research of Computers, 34(8): 2241-2246.
4. Zeru, L. (2022) Speech Recognition Method and Development Trend Based on Artificial Intelligence Deep Learning. NEW GENERATION OF INFORMATION TECHNOLOGY, 5(1): 104-106.
5. Zhenghong, WW. Su, P. Kun, Z. (2022) Research on Speech Recognition Based on Big Data and Deep Learning. Software, 43(01):133-135.
6. Rodríguez, ME. Ruíz-Mezcua, B. García-Crespo, A. et al. (1997) Speech/speaker recognition using a HMM/GMM hybrid model. Springer-Verlag, 1997:227-234.
7. Jinyin, C. Linhui, Y. Haibin, Z. et al. (2020) A black-box adversarial attack method for speech recognition systems. Small Microcomputer Systems,41(5):1019-1029.
8. Qingqing, Z. Yong, L. Zhichao, W. et al. (2014) The Application of Convolutional Neural Network in Speech Recognition. Internet New Media Technology, 3(6): 39-42.
9. Shouye, Z. (2022) Exploring End-to-End Deep Convolutional Neural Network Speech Recognition. Software, 43(3): 173-176.
10. Han, M. Roubing, T. Yi, Z. et al. (2022) Survey on Speech Recognition. Computer Systems & Applications, 31(1):1−10.
11. SHANGHAI JIAO TONG UNIVERSITY, 2019. Powerful end-to-end speech recognition framework-RNN-T. https://bat.sjtu.edu.cn/zh/rnn-t/