Research Progress Discussion on Deep Language Signal and Information Processing

Chunmei Wang
Linyi University School of Automation and Electrical Engineering, 276000, China

Abstract. This article focus on the progress discussion of deep speech signal and information processing. It will combine theory with practice, and deeply analyze the research progress of speech recognition technology, speech blending technique, and speech intensified technique. Based on the big data and cloud computing, the conclusion of valuable information can be effectively identified through deep speech signal and information processing technology. Hope it will help the relevant units.

1. Introduction
Based on cloud computing environment, a large amount of information and data are generated every day that including: text data, digital data, picture data, voice data, etc. Voice data is the most difficult to extract. It takes a series of proven techniques and methods to extract valuable information. However, the study on this aspect needs further research. Therefore, this article discusses the research process of deep speech signal and information processing based on the theory and practice.

2. Deep learning and deep neural networks
The concept of deep learning originates from the research of artificial neural network. Artificial neural network is a new model of machine learning and artificial intelligence in the new era. The concept of deep learning has laid a solid foundation for our country to simulate the development of neural networks. The perceptual layer belongs to the single-layer network structure, which has a very strong limitation in the processing of relevant data. After years of research, the multi-layer sensing layer is proposed, which is composed of multiple single-layer perceptrons. The main difference between the multilayer perceptron and the single-layer perceptron is that the multilayer perceptron is composed of multi-layer linear and its expressive ability is stronger. However, with the continuous development of voice information, the individual multi-layer perceptron has been unable to meet the actual demand. The concept of deep neural networks is to combine two or more multilayer perceptrons together. In the application process, the deep neural networks can be trained by the corresponding propagation algorithm. It mainly applies function model as nonlinear function. In general, the more hidden layers in the deep neural networks, the more obvious this phenomenon is. It is difficult for the deep neural network to exert its value and function. In 2006, experts and scholars like Hinton put forward a deep confidence network to solve this problem effectively.

3. Voice recognition technology
3.1. HMM-DNN acoustic model
Before 2006, the HMM -GMM acoustic model was the main speech recognition technology. But the model has strong limitations. It belongs to the medium shallow acoustic model, and it is difficult to meet
the specific requirements under the background of big data and cloud computing. After 2006, deep confidence network technology was developed. The HMM-DNN acoustic model can be developed with this technique. Compared with the traditional HMM-GMM acoustic model, the performance and other aspects have been greatly improved. The main advantages are mainly reflected in the following aspects:

Firstly, through deep neural network we can estimate posterior probability distribution ratio of HMM (Hidden Markov Model). There is no need to assume the voice distribution. It greatly improves the efficiency and accuracy of speech processing.

Secondly, in the process of application, deep neural network can be combined with many characteristics. It includes both discrete data and continuous feature data.

Thirdly, deep neural networks can also be extracted using the information contained in adjacent voice frames. In short, the deep neural network is used as a feature extraction network, which effectively improves the efficiency and accuracy of speech recognition.

In addition to the HMM-DNN acoustic model, CNN (deep convolutional network) and RNN (recurrent neural network) have also developed well in recent years. It is widely used in acoustic modeling of speech recognition. The acoustic model constructed by these two technologies can effectively extract the phonetic features unrelated to the peak frequency displacement of the speech spectrum. It can greatly improve the robustness of different speakers. RNN is more complex in application, and the corresponding technology is not mature enough. There is a high requirement for initialization, and training is time-consuming. The development has not yet been widely used in speech recognition of continuous features.

3.2. Training method of HMM-DNN acoustic model in big data environment

A large number of application examples show that, based on the HMM-DNN acoustic model, the comparison of the traditional HMM-GMM acoustic model is more significant in various aspects in continuous speech recognition. But deep neural network training is a very long and hard work. Even if the most advanced GPU is used for accelerated training, it will take three to six weeks to train a normal 6 hidden layer on a 1000h voice data set [1]. The main reason for the long training cycle is the basic algorithm used in the training of deep neural network. It is a stochastic gradient descent algorithm and the speed of convergence is slow. The essence belongs to a serial algorithm, it is difficult to carry on and deal with. Therefore, if the training cycle of neural network is shortened in the big data environment, it is one of the problems that should be solved in depth voice information processing.

There are two methods to shorten the training cycle of deep neural network. One is to improve the training efficiency by using the sparse simplified model structure of the deep neural network model parameters. The operating principle reduces the size of the model by forcing 80% of the smaller parameters to zero. However, the performance of the deep neural network is not lost, and the training efficiency can be increased by 30%~50%. The other is training by using two cpus. The application example shows that this method can improve the training efficiency of 3.1~4 times. But the cost is higher, and need more gradient information to realize it. It is also necessary to redistribute the relevant information and establish a new parameter model to achieve the accessibility of all data. When applying this technique, the data between different cell units is required to be transmitted frequently. It greatly increases the complexity of the application. Moreover, it requires the relevant experts and scholars to devote more energy to the research and realize the value and function of the application.

4. Research progress in speech synthesis technology

The speech synthesis technology based on HMM parameter is the most widely used speech synthesis technology. A large number of application examples show that compared with other technologies, the most obvious advantage of this technique is: the quality stability of synthesized speech is very high, and the requirements for storage and computing resources are low. In the specific application process, it can make scientific and reasonable adjustment to the tone color aspect. But it also has some disadvantages. For example, in the case of the original voice, the decline in sound quality is very obvious. The main reason is that the vocoder performance is not fully played, and the accuracy of acoustic modeling needs
to be further improved [2]. In order to address these shortcomings effectively, many experts and researchers have tried to introduce deep learning techniques into speech synthesis. The specific research progress is as follows:

Hnn-rbm and HMM -DBN speech synthesis technology is one of the most applied techniques in speech synthesis. However, in the process of application, it is necessary to make corresponding decision with the help of spectral parameters. In the specific synthesis process, the probability density function model of RBM or DBN is used to replace the Gaussian mean. Compared with other synthetic technologies, the main advantage of this technology is: it is possible to model the relativistic high Gaussian spectrum envelope directly to preserve the spectral details better. At the same time, RBM technology and DBN technology also have strong modeling capabilities. They can provide more intuitive and effective data and reference basis for the distribution of spectral envelope. Moreover, it can also be used to synthesize sound sources at the end of the channel (most vowels). The cascades are compatible with the acoustic theory of speech production and need not be adjusted for each filter. Based on this consideration, people combine the two and put forward the hybrid resonance peak model. We can replace the Gaussian mean of the traditional method with the predictive value of the deep neural network and then make the parameter synthesis. It can greatly improve the efficiency and quality of speech synthesis.

5. Speech enhancement technology

Speech enhancement technology is the main link in deep speech information processing. This technique has long been widely used in speech processing. Voice enhancement techniques can provide clearer and clearer data for voice processing. However, the research of this technology is not deep enough, and many technologies and methods are still simple in the experiment. But it has a very promising future.

The model of RBM and deep neural network in deep learning is good at processing and recognizing the structural information technology in data. It can obtain higher level structured information from the data underlying structured information, which greatly improves the efficiency and accuracy of relevant information. Therefore, it is necessary to integrate the RBM and deep neural network into the speech enhancement, which is the main content of the current in-depth language information processing research. The speech enhancement technique based on the ideal binary time-frequency masking estimation is proposed, as shown in figure 1:

![Figure 1. Speech enhancement technology based on regression deep neural network](image)

As can be seen from figure 1, this technology can greatly improve the intelligibility of speech by improving the non-stationary speech enhancement of low SNR. However, the loss of sound quality is serious and needs to be further studied. Several experiments show that multi-frame extension has great significance in improving the quality and continuity of language enhancement. To some extent, it indicates that the structural information of speech spectrum plays an important role in the increase of speech, and it is worthy of extensive application.

6. Deep voice information and processing information outlook

Through the above analysis, it can be seen that the HMM -DNN speech recognition acoustic model based on deep learning, the application performance of a large number of continuous speech recognition
can be improved by 30%~50% that compare with the traditional speech recognition acoustic model. The research of deep learning in speech recognition includes the following points:

Firstly, In the course of training, the BP algorithm with gradient descent is adopted. The parallel development of training is hindered to some extent. However, In the context of big data and cloud computing, there is a huge amount of data generated every day. This allows the deep neural network to undergo a longer cycle during the training process. It restricts the performance and value of deep speech information processing. In China, the research on deep voice information processing technology is late, and the technologies are not mature enough to solve the problem of the long time of network training. Therefore, it is necessary to pay more attention to the study more effective training methods in the later research.

Secondly, in the specific model structure, the speech recognition technology based on RNN acoustic model needs further research. RNN can directly model the timing of speech. Therefore, it has a very broad development space, and it is worth popularizing and applying.

Thirdly, Dnn-hmm is still the main research point in the next 10 to 20 years. The most effective technology at present is adaptive technology based on speaker coding. However, there are still a series of problems to be solved in the specific application process. For example, the code of the speaker does not have the physical significance of expressing the voice of the speaker. Therefore, the technology needs to be further improved [4].

Fourthly, deep learning technology in speech enhancement research include: further improve environment and is not included in the training set the noise influence the performance of speech enhancement. Thus, the adaptive problem of the deep neural network to the noise environment is greatly increased.

7. Summary
From the above, this paper deeply discusses the research progress of deep speech signal and information processing based on theoretical practice. The results show that the current national deep voice signal and information processing technology have achieved good development, but there are still some problems to be solved. This requires the researchers to invest more energy and further research in order to give full play to the value and role of deep learning in speech signals and information processing.

References
[1] Liu Wenju, Nie Shuai, Liang Shan, etc. Deep Learning Based Speech Separation Technology and Its Developments [J] Acta Automatica Sinica, 2016, 42(6):819-833.
[2] Hou Yimin, Zhou Huiqiong, Wang Zhengyi. Overview of speech recognition based on deep learning [J]. Application Research of Computers, 2017, 34(8):2241-2246.
[3] Dai Lirong, Zhang Shiliang. Deep Speech Signal and Information Processing: Research Progress and Prospect [J]. Journal of Data Acquisition & Processing, 2014, 29(2):171-179.
[4] Zhang Xiongwei, Li Yinan, Zheng Changyan, etc. Speech Dereverberation, Review of State-of-the-Arts and Prospects [J]. Journal of Data Acquisition & Processing, 2017, 32(6):1069-1081.
[5] LI Jing, Design of speech signal acquisition and processing system based on MATLAB, Journal of Shanxi Datong University(Natural Science Edition), 2016, 2,p.30-33
[6] Chen Pengzhan; Zhang Xin; Xu Fangping, Dual modal emotion recognition based on speech and text information, Journal of East China Jiaotong University, 2017, 2,p.100-104