Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/77/20240661

Application of deep learning-based speech signal processing technology in electronic communication

Yixuan Wu

Jiangsu Normal University Jiangsu Saint Polytech College - Sino-Russian Institute

1969162317@qq.com

Abstract. In recent years, the artificial intelligence boom triggered by deep learning is influencing and changing people's lifestyles. People are no longer satisfied with human-computer interaction through simple text commands; instead, they look forward to more convenient and faster communication methods like voice interaction. Against the backdrop of innovative development, the application of speech signal processing systems is becoming increasingly widespread. Therefore, it is necessary to study the application of deep learning-based speech signal processing technology in electronic communication. This can provide more valuable references and assistance for future development, promoting the better development of deep learning-based speech signal processing technology in electronic communication. In this paper, we first review the application of deep learning in speech signal enhancement, speech recognition, and speech synthesis from a theoretical analysis perspective. Then, we discuss the application of deep learning-based speech signal processing in electronic communication, including the application of models such as Transformer, LAS (Listen, Attend and Spell), and GFT-conformer in speech signal processing. We also discuss some application scenarios of deep learning-based speech signal processing in electronic communication. Finally, we identify the need for deeper application of deep learning technology in speech signal processing and electronic communication, with continuous optimization and adjustment.

Keywords. Deep Learning, Speech Enhancement, Speech Recognition, Speech Synthesis, Electronic Communication

1. Introduction

As a symbol capable of transmitting information, speech signals are well applied in various fields and industries. To some extent, their development has changed traditional information dissemination, playing a role in promoting and advancing the development of electronic information engineering.

With the continuous innovation of communication technology, speech has become an important medium for communication between humans and machines, as well as between machines themselves. Deep learning, as a powerful machine learning technology, brings new prospects to speech signal processing with its excellent performance in large-scale data processing and complex pattern recognition. Compared to traditional methods, deep learning models can better capture the complex features in speech signals, thus achieving remarkable results in tasks such as speech signal enhancement, speech recognition, and speech synthesis. This paper aims to explore the application of deep learning technology in speech signal processing and electronic communication. Based on the application of

DOI: 10.54254/2755-2721/77/20240661

different models, we discuss the issues and challenges of deep learning technology in speech signal processing to provide improvement directions for future research.

This paper is broadly divided into four parts. The second part of the paper will discuss the application of deep learning in speech signal processing. The third part of the paper will focus on the application of deep learning-based speech signal processing in electronic communication. Finally, the fourth part of the paper will provide a summary, discussing the limitations of this research and directions for future research.

2. Application of Deep Learning in Speech Signal Processing

2.1. Application of Deep Learning in Speech Signal Enhancement

With the development of modern technology, scientists have further integrated deep learning systems into the field of speech signal enhancement. Deep learning-based speech noise reduction algorithms, relying on their powerful learning and data processing capabilities, often demonstrate better performance and adaptability to changing acoustic scenarios compared to traditional speech noise reduction algorithms.

In 1979, Boll et al. first applied spectral subtraction to the field of speech noise reduction, proposing an efficient digital speech analysis algorithm independent of the processor [1]. The implementation of spectral subtraction treats the quiet periods of speech signals as noise. By averaging the intensity of the quiet periods, it estimates the noise signal and subsequently retrieves the estimated clean speech signal. Due to the averaging method, residual noise might remain after spectral subtraction. To improve this situation, many researchers have continuously studied and refined the algorithm. Berouti et al. added two parameters to the traditional spectral subtraction method [2], namely the over-subtraction factor and the spectral floor threshold. When the subtracted amplitude is less than the spectral floor threshold, the subtracted amplitude is set to the spectral floor threshold, reducing the impact of residual noise. Sarafnia et al. proposed a Kalman filtering algorithm based on Bayesian state-space [3], which uses the prior probability distributions of the target speech signal and noise signal to successfully recover clean speech signals from noisy speech signals. However, the large number of parameters and the complexity of deep learning models pose significant challenges for deployment on resource-constrained devices. Reference [4] proposed a deep learning model based on DNN, which connects the contextual information of speech signals to a feature vector for DNN learning, improving noise reduction performance in complex noise environments. Weninger et al. [5] proposed estimating clean speech and noise features in noisy speech by training a Long Short-Term Memory (LSTM) RNN, finding that using the Ideal Binary Mask (IBM) in Computational Auditory Scene Analysis (CASA) significantly improves speech intelligibility in noisy environments.

2.2. Application of Deep Learning in Speech Recognition

As an important interface for human-computer interaction, speech recognition has evolved significantly, bringing convenience to many aspects of our lives. However, its performance still lags behind human capabilities. The primary reasons are differences in speakers, environmental noise, and variability in recording devices. In recent years, due to the strong discriminative features extracted by deep learning and the strong discriminative ability of the trained models, researchers have applied deep learning to speech recognition. For instance, Mohamed et al. from the University of Toronto used a DBN network to build a monophone classifier [6]. Microsoft researchers, in collaboration with Hinton, used Deep Belief Networks (DBN) as a pre-training process for Deep Neural Networks (DNN), achieving great success in acoustic model training for large vocabulary speech recognition systems.

Traditional speech recognition systems require acoustic models, language models, and pronunciation dictionaries. The modernization of Tibetan information processing involves using traditional methods like DNN-HMM, which necessitate a pronunciation dictionary and a deep understanding of Tibetan, thus raising the entry barrier for Tibetan speech recognition processing. In recent years, the emergence

DOI: 10.54254/2755-2721/77/20240661

of end-to-end networks [7] has lowered the preparation threshold for speech recognition, consisting of an encoder and a decoder that only require speech and text for direct conversion from speech to text.

Moreover, end-to-end networks provide a broader choice of modeling units, leading to the proposal of various end-to-end models, such as Connectionist Temporal Classification (CTC) and attention-based LAS (Listen, Attend and Spell) models. Experts have built the Transformer model, using Positional Encoding and its core self-Attention mechanism for better speech recognition. In 2018, Dong et al. [8] first applied the Transformer model to ASR, while Huang et al. [9] studied streaming Transformer speech recognition networks by limiting the learning scope of self-attention. Streaming speech recognition involves real-time speech data input and real-time text output, requiring faster decoding speeds compared to non-streaming speech recognition (see Figure 1).

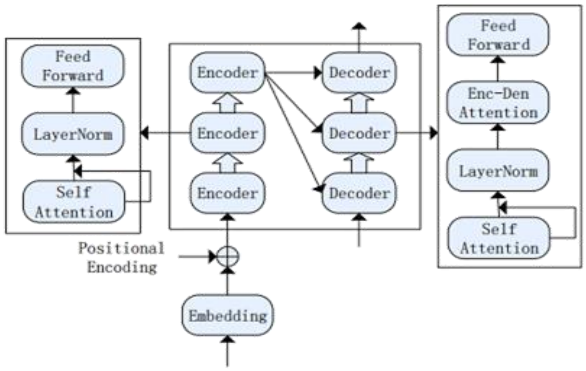

Figure 1. Transformer structure [10]

2.3. Application of Deep Learning in Speech Synthesis

With the development of economic globalization, communication between different languages is becoming increasingly frequent. Therefore, speech conversion has become a research hotspot in academia. Currently, speech synthesis systems using deep learning theories are gradually gaining popularity. More and more industry professionals and experts are building deep neural networks to establish a mapping relationship from text sequences to high-dimensional speech signals, comprehensively improving the performance of speech synthesis systems. Deep learning-based speech synthesis systems are mainly divided into two categories: one introduces deep learning technology into each module of traditional speech synthesis systems, and the other focuses solely on end-to-end speech synthesis systems, concentrating on the mapping from text sequences to intermediate representations such as Mel spectrograms and from intermediate representations to speech signals. Since Google's DeepMind team proposed Wavenet in 2016, end-to-end speech synthesis systems have seen rapid development. Reference [11] improved the Wavenet structure by directly using existing acoustic features for modeling, simplifying the modeling process, and enhancing model usability. The Char2Wav deep learning speech synthesis system consists of an encoder, decoder, and vocoder. Compared to Wavenet, the SampleRNN network structure is simpler, generating speech faster. However, because SampleRNN's underlying architecture is RNN, it cannot fully utilize computing resources and requires some acceleration techniques to speed up model training [12].

3. Application of Deep Learning-based Speech Signal Processing in Electronic Communication

Speech signal processing, as a product of the combination of phonetics and digital signal processing, is closely related to many disciplines. Experts have attempted to apply speech signal processing to the field of electronic communication, and significant advancements have been made in digital telephone communication, high-quality narrowband voice communication systems, and speech learning machines. Taking digital telephone communication as an example, China's current digital communication system has three main advantages: ①The transmission quality of communication systems is increasingly improving, and the quality of voice communication is continuously developing; ②The current digital

DOI: 10.54254/2755-2721/77/20240661

communication system has more external interfaces compared to the previous systems, enhancing the system's expandability; ③The current digital communication system has improved self-testing and intelligent alarm functions, reducing operating costs and enhancing system security [13]. With the integration of speech signal processing research, the development of electronic communication has been greatly enhanced.

3.1. Deep Learning Model Selection and Its Application in Speech Signal Processing

In modern automatic speech recognition and video conferencing applications, the perceived speech quality directly depends on the performance of the underlying speech enhancement system. Benefiting from the rapid development of deep learning, experts choose to use deep learning models for in-depth discussion in speech signal processing research. Currently, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are widely used in the field of speech enhancement and have achieved remarkable success. Reference [14] proposed a Convolutional Recurrent Network (CRN), which uses the magnitude spectrum of noisy speech as input, processes it through the network, reconstructs the speech spectrum using the original phase of the noisy speech, and obtains enhanced speech through inverse transformation. CRN combines the advantages of CNN and Long-Short Term Memory (LSTM) units, effectively utilizing temporal information. TAN et al. [15] improved CRN by proposing a Gated Convolutional Recurrent Network (GCRN), which decomposes the spectrum of noisy speech into real and imaginary parts, using two decoders to predict the real and imaginary parts of the speech spectrum separately. HU et al. [16], inspired by the Deep Complex U-Net, introduced complex operations into neural networks and proposed a Deep Complex Convolutional Recurrent Network (DCCRN). WESTHAUSEN et al. [17] proposed a Dual-Signal Transformation LSTM Network (DTLN), estimating the magnitude spectrum in the first stage and recovering the time-domain speech signal by combining the noisy speech phase. In the second stage, the time-domain signal is used as input to estimate the clean speech time-domain signal.

Describing speech enhancement as a supervised learning problem, noisy speech can be enhanced in the time domain or time-frequency domain using DNN. The time-frequency domain enhancement method [18] first uses Short-Time Fourier Transform (STFT) or other time-frequency analysis methods to convert speech signals into spectral representations, then denoises them in the time-frequency domain using DNN, obtaining enhanced speech signals. Compared to the transformer, the convolution-augmented transformer combines the characteristics of Convolutional Neural Networks (CNN) and transformers, effectively capturing both local and global dependencies of features, further improving the performance of time-domain and time-frequency domain speech enhancement methods.

3.2. Application Scenarios of Deep Learning-based Speech Signal Processing in Electronic Communication

With the continuous development of electronic communication technology, higher requirements are being placed on the security and reliability of electronic communication equipment. Applying deep learning technology to electronic communication has extensive practical application value in problems such as abnormal signal recognition in electronic communication devices. Taking speech recognition as an example, the automatic identification of abnormal signals in electronic communication equipment is based on feature analysis of the signals from these devices, constructing a multi-dimensional distributed sensing information tracking and recognition model to detect and extract the sensing information characteristics of electronic communication device signals [19]. The automatic identification method for abnormal signals in electronic communication devices, based on improved deep learning, first constructs a transmission channel equalization adjustment model for the signals of electronic communication devices. It uses deep learning methods to achieve output equalization scheduling of the signals, realizes signal matching and spectral analysis of electronic communication devices through beam interval equalization control methods, and extracts the features of electronic communication device signals by combining feature distributed extraction and parameter optimization estimation methods. Improved deep learning methods are used to extract features and achieve automatic

DOI: 10.54254/2755-2721/77/20240661

identification of abnormal signals in electronic communication devices. Since the deep learning neural network is formed by stacking autoencoders, the accuracy of the signal data of electronic communication devices obtained through such operations is relatively high [20]. Using the autoencoder of deep learning to perform stacked training on DNN, DNN can obtain the necessary abnormal signal feature information from the sample signal data provided by electronic communication devices.

4. Conclusion

This paper primarily provides a detailed examination of the application of deep learning-based speech signal processing technology in electronic communication. It introduces various aspects such as speech enhancement, speech recognition, and speech synthesis within speech signal recognition, focusing on speech signal preprocessing and feature extraction, model training and optimization, performance evaluation, and comparative analysis. The study aims to understand the current research status of speech signal processing in electronic communication and explore potential future research directions, trends, and applications in this field. Throughout the research and discussion process, we found that deep learning, as a powerful machine learning technology, offers new directions for speech signal processing with its excellent performance in large-scale data processing and complex pattern recognition.

Research on the application of deep learning-based speech signal processing technology in electronic communication has achieved significant results. These studies model, extract features, and reconstruct speech signals using deep learning models, effectively removing and suppressing interference signals such as noise and echo, thereby improving the quality and reliability of speech in communication systems. However, due to the limitations of technological development, deep learning in the field of speech signal processing still faces challenges, such as the impact of strong noise pollution. Based on these issues and challenges, future research can further explore the optimization of deep learning technology in speech signal processing to enhance the performance of electronic communication systems and provide better support for user experience.

References

[1] Boll S., Pulsipher D. Suppression of acoustic noise in speech using two microphone adaptive noise cancellation[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 28(6): 752-753.

[2] Berouti M., Schwartz R., Makhoul J. Enhancement of speech corrupted by acoustic noise[C]//1979 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Washington, USA: IEEE press, 1979: 208-211.

[3] Sarafnia A., Ghorshi S. Noise reduction of speech signal using bayesian state-space kalman filter[C]//2013 19th Asia-Pacific Conference on Communications (APCC). Denpasar, Indonesia: IEEE press, 2013: 545-549.

[4] Xu Yong, Du Jun, Dai Lirong, et al. An experimental study on speech enhancement based on deep neural networks[J]. IEEE Signal processing letters, 2013, 21(1): 65- 68.

[5] WENINGERF,EYBENF,SCHULLERB.Single --channel speech separation with memory--enhanced recurrent neural net--works{C}//IEEE International Conference on Acousties Speech and Signal Conference on Acoustics Speech and Signal Processing.Florence:IEEE Press,2014:3737--3741

[6] Mohamed Dahl,Geoffrey Hinton.Deep belief networks for phone recognition,Proc.NIPS Workshop,2009(12).

[7] LI J,WU Y,GAUL Y,et al. On the comparison of popular end-to-end models for large scale speech recognition [C] // Proceedings of the International Speech Communication Association,Interspeech 2020,Shanghai,China,2020:1-5.

[8] DONG L,XU S,XU B. Speech-transformer:ano-recurrence sequence-to-sequence model for speech recognition[C] // Proceedings of the 2018 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP),Calgary,Canada,2018: 5884-5888.

DOI: 10.54254/2755-2721/77/20240661

[9] HUANG W,HU W,YEUNG Y T,etal. Conv-transformer transducer:low latency,low frame rate,stream able end-to-end speech recognition[C] // Proceedings of the International Speech Communication Association, Inter speech 2020,Shanghai, China,2020:5001-5005.
[10] Gao, Y. R., & Bianba, W. D. (2023). Research on Tibetan speech recognition based on end-to-end deep learning. Modern Computer, 29(17), 25-30.
[11] Tamamori A, Hayashi T, Kobayashi K, et al. Speaker-dependent wavenet vocoder[C]. //Interspeech. 2017, 2017. 1118-1122.
[12] Shen J, Pang R, Weiss R J, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions[C]. //2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018. 4779-4783.
[13] Li, Q. (2019). Application of digital communication systems in railway transportation. Information and Communication, (07), 173+175.
[14] TAN Ke, WANG Deliang. A convolutional recurrent neural network for real-time speech enhancement[C]//Interspeech 2018. ISCA: ISCA, 2018: 3229-3233.
[15] TAN Ke, WANG Deliang. Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 380-390.
[16] HU Yanxin, LIU Yun, LV Shubo, et al. DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement[C]//Interspeech 2020. ISCA: ISCA, 2020: 2472-2476.
[17] WESTHAUSEN N L, MEYER B T. Dual-signal transformation LSTM network for real-time noise suppression[C]//Interspeech 2020. ISCA: ISCA, 2020: 2477-2481.
[18] LIN Ju, DE LIND VAN WIJNGAARDEN A J, WANG K C, et al. Speech enhancement using multi-stage self-attentive temporal convolutional networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3440-3450.
[19] Fu, S. (2019). Optimization scheme for abnormal uplink wireless communication signals on metering terminals. Electronic Production, (15), 93-94.
[20] Yan, W. L. (2021). Design of ship abnormal data communication recognition system for wireless local area networks. Ship Science and Technology, 43(4), 157-159.