International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

Research on Speech Separation Technology Based on Deep Learning Approach

Yue Hu, Ya Wang, Long Yang, Zhaolei Qi
School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu 233030, China

***Abstract*—** *Intelligent terminal is an emerging term, it is the product of the continuous development of Internet+ IT technology, in just a few years, various terminal applications supported by the mobile Internet have emerged one after another, subtly affecting people's daily life, has become an important part of people's happy life. With the convenience of intelligent applications and the characteristics of big data traffic, network data traffic has increased exponentially. Voice is one of the important carriers of language interaction, it is like a bridge to connec+t one side of the object and the other side of the object terminal, so more and more scholars begin to study intelligent speech, and the intelligent processing of speech has gradually become a hot spot in the market. As scholars introduce deep learning technology into the field of speech intelligent processing, it is possible for intelligent terminal services to process a large number of voices at the same time and can extract various effective information from input speech, such as emotions, meanings, keyword frequencies, and so on. However, in the process of practical application, speech is very susceptible to interference from various uncertain factors, there is "noise" in the process of voice propagation, which not only increases the difficulty of the work of the speech intelligent processing algorithm, but greatly reduces its performance, but also affects the accuracy of the processing results to a certain extent, which makes the entire speech intelligent processing process have great space for development and exploration, so the focus of scholars' research begins to gradually bias towards how to perform speech separation. Eliminate noise interference and achieve ideal application results in real-world scenarios. The research and development of speech separation technology is related to important applications in many fields, and speech separation plays the dual role of auxiliary tools and effective tools.*

***Keywords*—** *Speech Separation; Speech Intelligent Processing,; Real-world Scenarios.*

I. INTRODUCTION

In recent year, due to the advancement of science and technology, intelligent voice has been used in many fields, and its popularity in various fields is gradually increasing, and it is developing in the direction of diversification. With the continuous development of artificial intelligence technology, the intelligent voice industry will continue to move forward. In 2021, the State Council issued a number of policies on the construction of intelligent voice, which pointed out that all regions of the country need to strengthen intelligent applications such as intelligent voice and intelligent customer service to meet the individual needs of the masses. At the same time, the number of patent applications for intelligent voice has reached nearly 4,000.

Today, with the development of science and technology, the emergence of computers has made it possible for people to use computers to process speech. Computers can process repetitive events faster than the human brain, especially in the field of speech recognition. Voice plays a very important role as a bridge between people. Normal people can only convert speech into text to 60 words per minute, while computer speech recognition can reach 6,000 words per minute, and as the speed of the processor is gradually increasing, this speed is still growing. So far, voice is one of the important ways of interaction between people, and its convenience and directness are favored by the masses, so separating the target speech from the complex noise environment is the focus, hot spot, and difficulty of academic research.

For the speech separation system, the input is the user's recording or uploaded voice file, and the output is the result after voice separation. With the wide application of AI, speech separation brings great convenience, which plays an important role in promoting social stability and progress and national economic development, and its application in some special fields continues to benefit people, such as it can help hearing-impaired patients achieve intelligent "hearing remodeling" and help them create a better and happy life. For example, this technology also has important applications in national defence, which can obtain important information and intelligence for the military from a large amount of noise, and can reduce information differences, and avoid being in a disadvantageous position.

II. RELATED WORK

(1) Speech separation

The problem of speech separation stems from the famous "cocktail party effect"[1]. This topic is proposed for complex environments, wanting to obtain the logic and semantics of the speech you want to understand from constant noise interference and the speech output of others, and it is inseparable from people's attention to the special function of the auditory system—auditory choice [2]. What scholars expect is to actively eliminate noise interference through a scientific model, only listen to what they want to hear, and only understand the speech they want to understand. But often the computer model cannot achieve this effect, it will be affected by a variety of interference noise, so that the expected effect is not one, so in response to this phenomenon, it is urgent to carry out new exploration and research on speech separation technology, which is constantly combined with various computer technologies, from different angles, dimensions to establish a diversified speech separation model, through the model simulation of the auditory system, to

Yue Hu, Ya Wang, Long Yang, and Zhaolei Qi, "Research on Speech Separation Technology Based on Deep Learning," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 5, Issue 12, pp. 70-73, 2023.

achieve computer speech separation technology can be comparable to the human ear-brain processing mode, so as to be widely used in various areas of life.

With the development of computers and deep learning, people think of using convolutional neural networks to deal with speech enhancement problems, which has gradually become a new research trend in the field of speech recognition, and now there are many related methods proposed, after all, in the field of image and text processing, deep neural networks show a very strong application, able to deal with many things that were previously unimaginable [3].

For a long time, many scholars and scientists have invested a lot of time and energy in the field of speech. An important part of this is speech separation. Whether it's speech recognition, or voice interaction, you need to extract clear speech first, and then use speech recognition technology to convert speech into text, and finally hand it over to the computer to process the text. Voice interaction and speech-to-text have not been completely replaced by artificial keyboard typing and touch screen input, and a very important reason is that the effect of speech recognition is too low, and the accuracy rate needs to be strengthened.

(2) Convolutional Neural Network

Convolutional Neural Network (CNN) is a deep learning model widely used in images, audio, text, and other. Its name comes from the convolution operation, which is performed on one- or two-dimensional images to efficiently extract important information while preserving local features [4].

Speech separation based on CNN is one of the research directions that has received widespread attention in recent years. The main idea is to input the voice signal mixed by multiple sound sources into the CNN, and then learn the spatial and temporal domain characteristics of the sound source through the CNN model, and finally separate each individual sound source from the mixed signal. The following will introduce 4 popular CNN-based speech separation algorithms. For input data, such as images, CNN can gradually reduce the complexity of the data and extract features through multi-layer convolution and pooling. A typical CNN consists of several convolutional layers, pooling layers, fully connected layers, and activation functions [5]. The convolutional layer uses a set of fixed-size convolution kernels to process the data and operate on selected areas of the data. Between each layer, there is a trainable weight parameter that is gradually adjusted to optimize network performance based on the error of the loss function calculation.

In the speech separation tasks, CNN also have powerful expression and separation capabilities. The following describes the application of some CNN in speech separation tasks.

(a)Masking based on time frequency

Masking based on time frequency can be used as a simple but effective method for speech separation. Thus, more and more research is combining it with CNN to build a complete speech separation system. This type of method is often referred to as T-F masking. It uses phase and amplitude information from the spectrum and models different masking generators in the time-frequency domain. Some of the latest research have shown that this approach can achieve very good results.

(b)Techniques based on source activity

Techniques based on source activity refers to the use of source activity of speech signals to achieve speech separation tasks. This technique has proven superior in more complex speech scenarios because it uses dense sparsity coefficients to capture structural features inside speech signals from a higher dimensional perspective. In recent years, researchers have proposed some convolutional neural network structures for speech separation tasks based on source activity, such as sparse auto encoders and restricted Boltzmann machines.

(c)Based on depth inverse filtering

Based on depth inverse filtering refers to the use of CNN to estimate the spectrum of speech signals to achieve speech separation tasks. This method does not require the use of external masking information or speech corpus, so it tends to be more effective when the amount of data is small or the speech scene is more complex. Some recent studies have shown that convolutional neural networks based on deep inverse filtering technology can achieve better speech separation effects.

(d)Based on multi-scale convolution and attention mechanism

In recent years, a number of successful speech separation systems based on CNN have emerged, including various special architectures, such as multi-scale convolution and attention mechanisms, which can learn different levels of features to capture finer speech information.

In general, CNN shows very superior performance in speech separation tasks, so they are widely used in the field of speech separation. At the same time, new convolutional neural network structures are constantly emerging, which can also handle more complex scenario problems and further improve the quality and naturalness of speech signals.

(3) Translation invariance

For the basic subgraph pattern, these subgraphs can appear anywhere and have the same structure. Models that use translation invariance are particularly suitable for tasks that require a large number of local structures, thus providing critical correlation and sensitivity [6-7].

Overall, convolutional neural networks are a very efficient and flexible deep learning architecture that automatically extracts and learns complex features from data and uses them in a wide range of applications such as images, audio, text, and more. However, CNNs also have some limitations, such as large sample data requirements and uninterpretable internal features. But due to its powerful expressive ability and excellent performance, more and more tasks are using convolutional neural networks to obtain better model accuracy and more efficient results [8].

III. Application of Speech Separation

As an important speech signal processing technology, the application of speech separation technology is not only limited to the field of scientific research but also widely used in people's daily life and work. It plays an important role in

Yue Hu, Ya Wang, Long Yang, and Zhaolei Qi, "Research on Speech Separation Technology Based on Deep Learning," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 5, Issue 12, pp. 70-73, 2023.

various application scenarios such as military, medical health, and hearing aids, providing people with convenient services and meeting various needs.

(1) Military applications

Speech separation technology is widely used in the military field and can be used in real-time communication, intelligence monitoring, and UAV navigation and positioning. Especially in the combat environment, radio interference, confusing of thinking and noisy environmental sounds, and other factors often make the voice signal very confusing, and by using speech separation technology, the speech generated under these poor conditions can be effectively identified and separated, so as to ensure high-quality communication and traceability records and provide assistance for combat command and military monitoring.

(2) Hearing aids

Speech separation technology also plays an important role in hearing aid design. Many older adults and people with disabilities need hearing aids to improve their hearing. However, in noisy, multi-person meetings, public settings, etc., hearing aids may not be effective in separating and recognizing sounds. By using speech separation technology, existing hearing aids can keep background noise and other people's speech away from the listener's ear, making it easier for them to hear the sound signals that they care about.

(3) Healthcare

The application of speech separation technology is also very common in the medical and health field. For example, when drug advertisements or patient suggestion input, the quality of communication between patients and medical staff is improved by using voice separation technology to make real-time text communication more convenient to convert to text communication on the screen. In addition, on the other hand, speech separation technology is also used in this field to accurately grasp the language information of the patient's specific state for timely diagnosis and treatment, thereby reducing the incidence of medical errors and improving the depth and efficiency of rehabilitation.

(4) Internet applications

Speech separation technology has been widely used in Internet applications, such as video voice chat, virtual emotional communication, and other fields. This technique helps users separate background noise from other people's voices so that it is not annoying when chatting with multiple people or in groups. In addition, at the same time, speech separation technology can also reduce problems such as communication cards and high phone bills through the processing process.

On the whole, speech separation technology, as an important voice signal processing technology, plays an important role in military, medical, Internet applications and other fields. For example, there are a large number of practical applications in remote meetings, educational presentations, multi-person smart homes, social networking platforms and many other aspects. In the future, with the continuous advancement and development of technology, it is believed that this technology will have a wider range of application scenarios, bringing more convenience and innovation to people's life and work.

IV. Conclusions and Future Works

The rise of deep learning has opened up new research directions for speech separation [9-16]. It is hoped that in the future, technology can be taken to the next level and can exert greater value. With the advancement of science and technology, in the future society, the application of artificial intelligence and human-computer interaction has become popular is just around the corner, and its application will run through people's daily life, while driving the innovation and development of related industries, thereby bringing more benefits to the people, bringing more power to the society, and bringing more economic benefits to the country. At that time, people will live in a different, new era.

Currently, the research of speech separation has been very popular, and with the unremitting efforts of scholars at home and abroad, speech separation technology has entered a new level [10-19]. One of the most significant is supervised speech separation, scholars have carried out empirical research and algorithm improvement from different angles, and the most important has reached agreement on its essential characteristics and goals, and the various speech separation models derived from this tend to mature, forming a general speech separation model architecture, under this large architecture, extending in different directions, can map different directions, and be applied in different fields. However, it also exposed an obvious problem, that is, once a large structure is formed, it has relative stability, and it is undoubtedly difficult to break it to achieve a new round of innovation and development. For the speech separation architecture that supports the existing research results, we believe that speech separation will tend to develop in the following aspects:

(1) Constant generalization. Although current research has basically realized speech separation in various situations, after the introduction of deep learning, supervised speech separation operations can also be realized to a large extent. However, there are also great limitations, due to the limited coverage of the data, the full range cannot be achieved, which will lead to if the data cannot be matched, the results of speech separation will be unreliable, and the real voice separation needs cannot be realized. If data coverage is carried out continuously, the cost is too high and the method is single, which is not conducive to flexibility. We believe that speech separation technology can be generalized in the following two ways: One is analyze patterns in the ear's auditory system, combining CASA and speech separation; The other is the scope and focus of the study can be broadened, and due to the uncertainty of noise, it cannot be distinguished by a large feature framework, so we can change the direction and explore more of the essential and inherent characteristics of speech.

(2) Generative and supervised mode fusion. Speech separation technology is computer-implemented, but computers cannot have advanced modes like the human brain to control the

Yue Hu, Ya Wang, Long Yang, and Zhaolei Qi, "Research on Speech Separation Technology Based on Deep Learning," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 5, Issue 12, pp. 70-73, 2023.

processing of speech by the human ear auditory system. So we can continue to explore the speech processing paradigm in the human brain to adjust the computer simulation. The main thing among them is the unit of understanding of speech, the human brain can achieve fast and accurate response because the human brain has a complete set of mature modes of storage and combination, and matching, which can help us extract effective speech signals and parse them in various interferences. Computer speech separation technology can also be similarly combined with a generative system and training system, which may bring new breakthroughs to speech separation technology.

ACKNOWLEDGMENT

We thank the anonymous reviewers and editors for their very constructive comments. This work was supported in part by the Undergraduate Research and Innovation Fund project of Anhui University of Finance and Economics under Grant No. XSKY23157.

REFERENCES

[1] Ephral A,Mosseri I,Lang O,et al.Looking to listen at the cocktail party:a speaker-independent audio-visual model for speech separation [J]. ACM Transactions on Graphics,2018,37(4):1-11.
[2] Cherry E C. Some experiments on the recognition of speech, with one and with two ears[J]. The Journal of the acoustical society of America, 1953,25(5):975-979.
[3] Zmolikova K,Delcroix M,Kinoshita K,et al. Speaker Beam: Speaker aware neural net workfor target speaker extraction in speech mixtures[J].IEEE Journal of Selected Topics in Signal Processing,2019,13(4):800–814.
[4] Fan C H,Liu B,Tao J,et al.An End-to-End Speech Separation Method Based on Conv olutional Neural Network[J]. Journal of Signal Processing,2019,35(4):542-548 ·
[5] Li L,Girin L,Gannot S,Horaud R .Multichannel speech separation and en hancement using the convolutive transfer function[J].IEEE/ACM Transactions on Audio, Speech and Language Processing,2019,27(3):645–659.
[6] Kwon K,Shin J.Kim N S. NMF-Based Speech Enhancement Using Bases Update[J]. IEEE Signal Processing Letters, 2015, 22(4): 450-454.
[7] Erdogan H, Hershey J,Watanabe S,et al.Phase-sensitive and recognition boosted speech separation using deep recurrent neural networks[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
[8] Kolbæk M,Yu D,Tan Z-H,et al.Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(10): 1901-1913
[9] Wang Y,Wang D.A deep neural network for time-domain signal reconstruction[C].IEEEInternational Conference on Acoustics,Speech and Signal Processing(ICASSP).2015:4390–4394.
[10] Liu Y,Wang D.Divide and conquer:A deep CASA approach to talker-independent monaural speaker separation[J].IEEE/ACM Transactions on audio,speech,and lan guage processing,2019,27(12):2092–2102.
[11] Luo Y,Mesgarani N.Tasnet:Time-domain audio separation network for real-time, single-channel speech separation[C].IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP).2018:696–700.
[12] Boll S.Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1979, 27(2): 113-120.
[13] Han K,Wang Y,Wang D.Learning spectral mapping for speech dereverberation and denoising[J].IEEE/ACM Transactions on Audio,Speech and Language Processing, 2015,23(6):982–992.
[14] Chen Jingdong, Benesty J,Huang Yiteng, et al. New insights into the noise reduction Wiener filter[J]. IEEE Transactions on audio, speech, and language processing, 2006, 14(4): 1218-1234.
[15] Wang Q,Muckenhirn H,Wilson K,et al.Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking[J].IEEE/ACM Transactions on Audio,Speech,and Language Processing,2021,29:2840–2849.
[16] Li Y,Liu J,Liu Y,et al.Deep transduction non-negative Matrix decomposition Parallel Algorithm for Speech Separation [J]. Computer Science, 2020,47(8):49-55.
[17] Stoller D, Wang Y X, He L,et al. Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation [J]. IEEE Journal of Selected Topics in Signal Processing, 2018, 12(5): 1143-1159.
[18] Cakir E, Heittola T, Erdogan H, et al. Conv-TasNet: Surpassing Ideal Time-Frequency Masking for Speech Separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(8), 1256-1266.
[19] Ezambaygi S H, Samiee K, Mousavi H. D3Net: Separating Speech from Interferences by Recursive Parallel Interlace Convolution [J]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2021: 463-464.

Yue Hu, Ya Wang, Long Yang, and Zhaolei Qi, “Research on Speech Separation Technology Based on Deep Learning,” *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 5, Issue 12, pp. 70-73, 2023.