Research on the Application of Visual Maps Based on CiteSpace in Speech Technology

Jiake Yin
Chongqing University of Posts and Telecommunications, Chongqing 400065, China

*Corresponding author: 2018215030@stu.cqupt.edu.cn

Abstract. In recent years, deep learning has been widely used in the field of speech technology, and provides people with an efficient and convenient way of obtaining information and communicating in most application scenarios. This paper uses CiteSpace information visualization software to visually analyze the speech technology research literature based on nearly 10,000 papers in the field of speech technology research and application from 2013 to 2021. From the perspective of bibliometrics, this paper analyzes the distribution of hotspots and research frontiers in speech technology research in countries, institutions, disciplines and other information visualization maps, and compares and analyzes the literature information in the field of speech technology research and application at home and abroad in recent years. The research results show that the main disadvantages of speech technology at home and abroad are: the inability to completely avoid noise interference; the inability to correctly identify ambiguous and generalized sentences; and the inability to idealize the radio remotely. Aiming at the problem of noise interference, this paper proposes a deep learning-based noise reduction algorithm and increases the number of convolutions of the neural network to improve the model; then, for the ambiguity and generalization in the recognition process, this paper proposes a deep learning-based unsupervised training method; Finally, for the problems of sound reception and echo, this paper proposes a far-field sound pickup technology based on multi-channel signal microphone collection.

1. Introduction
Speech technology is the realization of human-machine language communication, including speech recognition technology (ASR) and speech synthesis technology (TTS) [1]. As early as the 1950s, Davis and others of AT&Tbell Labs successfully developed the world's first Audyr experimental system that can recognize the pronunciation of ten English numbers. In the 1970s, Soviet scientists first proposed a dynamic programming method to solve the problem of unequal speech signal length, and on this basis developed a dynamic time warping DTW algorithm [2]. At present, we use more Google Voice, Amazon Alexa, Microsoft Cortana and Apple's Siri, etc., all of which are perfect products based on this system. The research on speech technology in China began in the 1980s. The development strength of companies such as Huawei and iFlytek is comparable to that of foreign countries. The system recognition rate can reach up to 90%, the tone recognition rate is 99.5%, and the vocabulary recognition rate is 95% [3]. Since the 1990s, artificial neural network technology has been used as an entry point for speech recognition, making speech recognition from theory to practical application [4].
With the rapid development of the Internet and artificial intelligence technology, intelligent speech technology has become the most convenient and effective means for people to obtain information and communicate, and its development and advancement have also made great breakthroughs. In the 1980s, the establishment of the hidden Markov model made a great contribution to the development of speech technology. Since 2000, human-computer speech interaction has developed into a hotspot in speech technology research. In 2009, deep learning was applied to the field of speech technology. Compared with the traditional Gaussian mixture model-hidden Markov model (GMM-HMM), the speech recognition system achieved more than 20% relative sexual improvement [5]. In 2011, the application of deep neural network (DNN) has greatly improved the continuous speech recognition technology of large vocabulary, and achieved the biggest breakthrough in speech recognition effect in the past 10 years. Since then, the deep neural network (DNN)-based modeling method gradually replaced the GMM-HMM and became the mainstream of speech recognition modeling [6], which greatly promoted the development of intelligent speech technology. The development has broken through the bottleneck of intelligent speech technology in actual application scenarios, making the speech recognition technology truly practical. On this basis, speech recognition relies on the continuous development and innovation of big data technology and cloud computing technology [7].

Currently, speech technology has been widely used in various scenarios. Voice assistants in the form of mobile applications include Apple's Siri, Amazon's Alexa, Google Assistant, and Microsoft's Cortana. Smart speaker products include Amazon's Echo, Google's Home, and Apple's Home [8]. In daily life, speech technology can help people turn on navigation, play music, send text messages, make calls, control smart home devices, order food, order tickets, check information, and so on. In the medical field, environmental listening systems try to simulate human medical staff by recognizing the dialogue between clinicians and patients and capturing clinically relevant information [9]. No matter what kind of application, speech technology is changing people's lifestyles and habits, and has become an important part of human social life. However, noise interference still exists in the application process, generalization and ambiguity of vocabulary and sentences, poor long-distance reception, and reverberation. This article aims to explore the best solutions to the shortcomings of speech technology in the application process by studying the development fields, development hotspots and progress of speech technology in recent years.

2. Analysis methods and data collection

2.1. Software introduction

CiteSpace is a bibliometric software for batch research literature. The main principle is to use the visualization of knowledge graphs, and combine mathematics and statistics methods to quantitatively analyze the knowledge carrier. According to the author, institution, country, keywords and other information of the literature, a visual knowledge map is automatically generated to analyze the research status, hotspots and frontier dynamics of the field [10], which can be used to summarize the development process and shortcomings of the research field and explore future popular applications and development trends, etc.

2.2. Data collection

The data analyzed in this paper comes from the Web of Science core collection database in Web of Science. Log in to the Web of Science homepage, and enter the combination of "speech technology", "voice technology" and "speech recognition" in the basic search to search for all fields with unlimited dates, journals, and languages. By the time of data collection (17 March 2021), Web of Science Party included 1985–2021 speech technology-related research literature up to 20,175. In order to obtain the latest research hotspots of speech technology, the first 10,000 records were selected for analysis, of which 9,925 were valid documents, and the time span was from July 2014 to April 2021. After marking the retrieved data, select "Export to other file format", select the record content as "Full records and
2.3. Data analysis
Import the data into Citespace, set the time division from January 2014 to April 2021, set the time slice (Year Per Slice) to 1 year, and select the term source as Title, Abstract, Author and Key Word, Keyword Plus. Check the countries, institutions, keywords, cited documents, etc. respectively. The algorithm analysis part is composed of Pathfinder and Pruning slices. Visualization section check "Cluster View" and "Show Merced Network". By selecting the top 30 projects with the most cited or most occurrences from each slice, a visual knowledge map of speech technology research literature can be generated, and then the current research status and frontiers in the field of speech technology can be analyzed through the map.

3. Analysis of academic cooperation

3.1. Map analysis
In order to deeply analyze the academic cooperation in the field of speech technology, the node types of CiteSpace are respectively set as countries and research institutions, and the top 30 projects with the most occurrences are selected from each slice to generate a clustering map of research countries and research institutions. Perform metrological analysis on the countries and institutions involved in the map to obtain research cooperation in the field of speech technology. The size of the nodes in the figure represents the amount of documents issued by each institution in the field of speech technology. The connections between the nodes indicate the degree of scientific research connections between various countries and institutions. The thicker the connection, the closer the cooperation relationship.

According to the cluster diagram 1, we can see that the research of speech technology is widely distributed in the world. The red line frame of Figure 1 shows that countries led by the United States, China, Australia, Canada, etc. have launched corresponding research in the field of speech technology and have had a greater impact. Among them, the United States and China have the most extensive research on speech technology. In addition, the United States, China, India, Japan and other countries have less cooperation. On the contrary, the ties between EU countries are relatively close. The main reason for this phenomenon is that the United States, China, India, Japan and other countries have developed rapidly in the field of technological innovation and artificial intelligence, and have the ability to independently research. The development of speech technology among European countries is inseparable from the support of the EU’s relevant technical policies, making EU countries close, mutually beneficial and win-win.

From the cluster diagram 2, it can be seen that the Chinese Academy of Sciences, Tsinghua University, Johns Hawkins University, Carnegie Mellon University and other prestigious institutions have made rapid progress and achieved good research results. The connections between nodes of research institutions in different countries are relatively sparse, and there is a certain degree of cooperation between institutions in the same country. The Chinese Academy of Sciences, Tsinghua University, and Johns Hawkins University in the United States have communication and exchanges in the field of speech technology. Shanghai Jiaotong University and the University of Cambridge in the United Kingdom also have a related cooperative relationship in the field of speech technology.
Figure 1. National distribution map of speech technology literature research

Figure 2. Co-occurrence map of speech technology literature research institutions
3.2. Analysis of cooperation

The research of speech technology in various countries is inseparable from all aspects of support. In terms of policy, the White House of the United States has successively issued several government reports on artificial intelligence, and countries such as the United Kingdom, the European Union, and Japan have issued artificial intelligence-related strategies and action plans, focusing on building artificial intelligence first-mover advantages. At the same time, my country has also promoted the development of artificial intelligence to the national strategic level, and increased policy support for the research and development and industrialization of intelligent speech technology [11]. In terms of technology, the development of artificial intelligence and the wide application of deep learning have laid a good foundation for the development of speech technology.

Table 1. Publication of national literature on speech technology research

| NO. | Country         | Number of Posts (pieces) | Percentage (%) |
|-----|----------------|--------------------------|----------------|
| 1   | United States  | 2861                     | 28.83          |
| 2   | China          | 1605                     | 16.17          |
| 3   | India          | 896                      | 9.03           |
| 4   | Japan          | 672                      | 6.77           |
| 5   | Germany        | 599                      | 6.04           |
| 6   | United Kingdom | 539                      | 5.43           |
| 7   | South Korea    | 367                      | 3.70           |
| 8   | France         | 333                      | 3.36           |
| 9   | Canada         | 257                      | 2.59           |
| 10  | Taiwan, China  | 216                      | 2.18           |

Judging from the number of documents published in various countries from July 2014 to April 2021 (Table 1), the United States published 2861 articles, accounting for 28.83%, China published 1605 articles, accounting for 16.17%, and India published 896 articles. Accounted for 9.03%. The top three countries in the number of research articles published a total of 5,362 articles, accounting for more than 50%. In addition, the number of research documents in Japan, Germany, and the United Kingdom has exceeded 500, which has made a great contribution to the development of speech technology.

In 2007, the Cambridge University team founded AI Speech, aiming to make human-computer interaction more natural and convenient. In 2008, AI Speech Co. Ltd. settled in Suzhou Industrial Park, officially started the domestic industrialization of speech technology, and began the exploration of speech technology in China. The full-link intelligent speech interaction technology independently developed by AI Speech has promoted it to become a highly recognized speech technology company. In 2014, the "3rd China-Canada (Ontario) Technology Transfer and R&D Cooperation Forum" was held in Toronto. Canada and China aim to give full play to their respective advantages and achieve complementary advantages in technology. In 2019, iFLYTEK signed a contract with Deloitte Australia. The two companies will cooperate in language transcription and translation, and use intelligent speech technology in Australian medical institutions. In 2020, Google expanded the Duplex AI phone speech system to the United Kingdom, Canada and Australia, breaking through the bottleneck of the human-machine dialogue system that machines cannot understand humans, and strengthening the cooperative relationship between countries.
Table 2: Frequency of Papers Published by Research Institutions of Speech Technology Literature

| Number | Institution                  | Frequency |
|--------|------------------------------|-----------|
| 1      | Chinese Acad Sci             | 175       |
| 2      | Johns Hopkins Univ           | 155       |
| 3      | Google Inc                   | 140       |
| 4      | Carnegie Mellon Univ         | 128       |
| 5      | Tsinghua Univ                | 121       |
| 6      | Univ Cambridge               | 119       |
| 7      | NTT Corp                     | 113       |
| 8      | Shanghai Jiao Tong Univ      | 102       |
| 9      | Univ Sci & Technol China     | 93        |
| 10     | Univ Sheffield               | 92        |

As can be seen from Table 2, since 2014, The Chinese Academy of Sciences, Tsinghua University, Shanghai Jiao Tong University, University of Science and Technology of China, John Hawkins University, Google Corporation, Carnegie Mellon University, Cambridge University, Sheffield University and Japan Telecom Company have made a high contribution to the research of speech technology.

Since the 1960s, Reddy and others of Carnegie Mellon University in the United States have carried out continuous speech recognition research [12], which has made great contributions to the development of speech technology in the United States and the world. In 2007, the "Speech and Language Technology Research Center of Tsinghua University Information Technology Research Institute" was formally established, and under the leadership of relevant research team members, it began to develop technologies and applications with independent intellectual property rights. In the same year, Google acquired an Internet phone company and began to handle Google voice services. From 2010 to 2011, the "Chinese Academy of Sciences-Key Laboratory of Language Acoustics and Content Understanding" was approved to be upgraded to the Chinese Academy of Sciences and Beijing Municipal Key Laboratory, which carried out more in-depth research on speech analysis and processing. In September 2011, the "National Engineering Laboratory for Speech and Language Information Processing" was officially inaugurated. This laboratory relies on the iFLYTEK Co. Ltd. and the University of Science and Technology of China to build the Speech Synthesis Laboratory, the Speech Recognition Laboratory, the Natural Language Processing Laboratory, and the Intelligent Human-Machine Voice Four core technology research laboratories, including the Interaction Research Center, continuously improve the research quality in the field of speech technology research. In 2012, Shanghai Jiao Tong University established the "Shanghai Jiao Tong University-AI Speech Laboratory", under the leadership of Yu Kai, Ph.D. graduated from Cambridge University, dedicated to the research of intelligent speech in human-computer interaction. In 2019, the team of the Key Laboratory of Language Acoustics and Content Understanding of the Chinese Academy of Sciences won the first place in the DCASE2019 Audio Scene Competition. In terms of audio scene recognition technology under matching equipment, team members explored the use of a variety of long and short-term features, and combined with deep learning-based data enhancement methods to achieve a test accuracy of 85.2%, which is a significant lead of 1.4% in the second place, and it far exceeds the discrimination ability of human beings [13]. Under the leadership of these research institutions supported by the state, government and speech technology experts, speech technology research has gradually moved from the initial "problem-laden" to intelligent, and has strengthened the research and development of core technologies in the process of continuous development.
4. Analysis of research fields

4.1. Map analysis
Change the node type to the category and select the top 30 items with the most occurrences to obtain the distribution map of the subject field as shown in Figure 3. It can be seen from the red line frame in the figure that computer science, engineering, electrical and electronic engineering, artificial intelligence, acoustics and other disciplines are the more cutting-edge disciplines in the field of intelligent speech technology research, and more in-depth experiments have been carried out in the development of speech technology. The thick and dense connections between nodes in various disciplines indicate that related disciplines cooperate closely in the field of speech technology research. Multi-scenario applications of speech technology cannot be realized by only one discipline, and multidisciplinary collaboration and joint innovation are required.

For example, in modern computer science, the technology that allows computers to understand people's speech is called computer speech automatic recognition technology, and the technology that enables computers to output human speech is called computer speech synthesis technology [14]. Automatic speech recognition technology (ASR) and speech synthesis technology (TTS) are two key technologies in the field of speech technology, so it can be said that speech technology is an important application in the field of computer science. With the development of the subject of artificial intelligence, the speech recognition system has been better trained and can perform intelligent processing more accurately and quickly. In addition, the processing of speech signals is closely related to the disciplines of electrical and electronic engineering, and speech recognition technology is also inseparable from the construction of acoustic models.

Figure 3. Discipline distribution map of speech technology literature research
4.2. Academic impact analysis

In order to analyze the academic influence of the clustering map of the speech technology research field, the top ten subjects of the speech technology research discipline are counted in Table 3. We call the node that connects several articles at the same time as the "intermediary". The sparseness of the connections around the node represents the centrality of the intermediary. Betweenness centrality is a measure of the centrality of nodes in the network, and it can also be used as a scientific measurement index for quantitative analysis [15]. It is equal to the number of shortest paths from all vertices to all other vertices passing through the node. From the perspective of scientific metrics, documents with high betweenness centrality usually play a very important role in the process of information transmission [16].

Table 3. Distribution of speech technology research disciplines

| Ranking | Discipline                        | Centrality | Frequency |
|---------|-----------------------------------|------------|-----------|
| 1       | Computer Science                  | 0.52       | 5072      |
| 2       | Engineering                       | 0.71       | 4774      |
| 3       | Electrical & Electronic Engineering | 0.31       | 4417      |
| 4       | Artificial Intelligence           | 0.13       | 2780      |
| 5       | Acoustics                         | 0.09       | 1816      |
| 6       | Theory & Methods                  | 0.08       | 1697      |
| 7       | Information System                | 0.02       | 1050      |
| 8       | Interdisciplinary application      | 0.40       | 882       |
| 9       | Otolaryngology                    | 0.07       | 875       |
| 10      | Audiology & speech pathology      | 0.13       | 831       |

The combination of speech technology and computer science becomes computer speech integration technology (CTI technology), which covers multiple aspects of computer and speech communication technology, and has been successfully applied to interactive speech response technology [17]. The advent of this technology has opened up a new era in the collaborative development of speech technology and multidisciplinary fields, leading speech technology to achieve remarkable results in multidisciplinary application scenarios.

It can be seen from Table 3 that engineering has the highest integration in the interdisciplinary research of speech technology research. Engineering is the process of using the principles of mathematics, physics, and other natural sciences to design useful objects. It is a master of many disciplines such as computer science, electrical and electronic engineering, and artificial intelligence. The realization of the application level of speech technology needs to closely link various disciplines to provide a disciplinary foundation for the research of speech technology in different applications and different technological innovations.

For example, intelligent recognition systems such as voice assistants need to establish acoustic models and train them. After receiving speech information, they need to perform signal processing and pattern recognition. These are related to acoustics, electronics and electrical engineering, computer science, artificial intelligence and other disciplines. Speech synthesis technology is also based on the establishment of acoustic models. It synthesizes intelligent speech through multiple technologies such as natural language understanding and digital signal processing and outputs it to the demander. It is the product of the combination of acoustics, linguistics, computer science and other disciplines.

With the development and improvement of speech technology, application scenarios such as daily assistants, smart homes, and smart medical care have gradually become the main development direction of speech technology. In addition, the recognition rate of cochlear implants is getting higher and higher, bringing convenience to the life of deaf-mute people, which has also become a major breakthrough in speech technology. Therefore, in recent years, speech technology has also been deeply studied in the disciplines of acoustics, otolaryngology, audiology and speech pathology.
5. Research hotspots and frontier analysis

5.1. Keyword map analysis

Keywords are the key content of a literature research, symbolizing the research hotspots, frontiers and research preferences in this field. In bibliometrics, high-frequency keywords are generally used to characterize research hotspots in this field. The higher the frequency of a keyword, the higher its attention in this field. Keyword co-occurrence, that is, when the same keyword appears in different documents, these documents will be connected by a line. It is a manifestation of the relevance of the thesis and represents the relevance of various disciplines in the field [18].

Set the node type of CiteSpace as keywords, and select the top 30 items with the most occurrences from each slice to generate a keyword co-occurrence map (Figure 4). It can be seen from the figure that the hotspots in the field of speech technology are closely related, and the core technologies of each application module are closely following the development trend, and have relatively similar unresolved deficiencies. With "speech recognition" as the center, keywords such as "recognition", "neural network", and "noise" are closely related to it. At present, in the field of speech technology development, the research content is relatively one-sided and the problems to be solved are relatively concentrated, and a certain degree of innovation and pioneering is still needed.

![Figure 4. Keyword co-occurrence map of research on speech technology literature](image-url)

Sort the occurrence frequency of keywords, and count the top 10 keywords with the highest frequency to get Table 4. According to Figure 4 and Table 4, it can be seen that the connections between the nodes are messy and concentrated in "speech recognition", "neural network" and their surrounding
areas. In recent years, the key area of speech technology research is "speech recognition", that is, speech recognition technology. As the core mechanism of speech technology, the main methods have experienced the development and update of "deep neural network" to "convolutional neural network" and "recurrent neural network", but these methods are all based on "neural network" training method. The problems that need to be solved are focused on "perception" cognitive ability, "noise" interference and so on.

The high-frequency vocabulary presented above are all developed around speech technology, representing key parts of the field's application scope, technical solutions, research hotspots, and future prospects. These keywords are inextricably linked, interlocking and indispensable, showing the complete system of the development of speech technology in recent years.

| No. | Words                        | Frequency |
|-----|------------------------------|-----------|
| 1   | Speech recognition           | 2633      |
| 2   | Automatic speech recognition | 621       |
| 3   | Recognition                  | 569       |
| 4   | Deep neural network          | 536       |
| 5   | Noise                        | 528       |
| 6   | Neural network               | 522       |
| 7   | Deep learning                | 463       |
| 8   | Perception                   | 449       |
| 9   | Speech                       | 418       |
| 10  | Cochlear implant             | 402       |

From an overall point of view, speech recognition technology is at the core of the research of intelligent speech technology, and it is inferred that in the future development, intelligent speech recognition will be in a more cutting-edge development direction. Smart applications such as Siri, Xiaodu smart speakers, Huawei Xiaoyi voice assistant, and AutoNavi Map voice assistant are all products of the field of speech recognition technology. During the recognition process, environmental noise interference and dialect speech recognition rate often occur. Not high, fuzzy far-field speech recognition [19] and other defects. These issues have become the next stage of research hotspots in various countries and disciplines, and also symbolize the future key application prospects in the field of speech technology. In addition, the overall co-occurrence network shown in the figure is relatively sparse and has few research hotspots. This shows that the research of speech technology is not mature enough, the research direction is limited and scattered, the research results that have been made are relatively one-sided, and there is a lack of innovation in combination with more disciplines.

5.2. Analysis of the evolution of hot frontiers

On the basis of the hot key words of voice technology, select "TimeLine view" to obtain the distribution of research hotspots and research topics of voice technology in different years, and get the evolution trend of voice technology research field (Figure 5). Analyzing the keywords and time in the picture, the voice technology developed rapidly in 2014 and developed in a broad direction, and gradually stabilized in 2016. The emergence of the "end-to-end" recognition system in 2018-2019 has made the accuracy of semantic understanding reach a level that is basically usable in some vertical fields [20]. At present, the development of voice technology will not only be used in scientific research scenarios, but will also be integrated with life to provide humans with high-quality, high-happiness intelligent services.
Combining the application situation and development characteristics of speech technology from 2014 to 2021, the evolution analysis shown in Table 5 is obtained.

**Table 5. The evolution of voice technology**

| Time     | Keywords                                                                 | Development and characteristics                                                                                                                                 |
|----------|--------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 2014-2015| "Cochlear implant", "perception", "noise", "automatic speech recognition technology", "normal hearing", "elderly", "neural network", "deep learning", "mel cepstrum coefficient MFCC", "feature extraction", and "recurrent neural network" | Cochlear implanters often have difficulty understanding noise interference and voices from a distance. The use of assistive hearing technology (HAT) has effectively improved the above problems [21], and has brought convenience to the deaf-mute group and the elderly with hearing loss. The MFCC algorithm is mainly used for feature extraction to extract the features of the speech signal, and the DTW algorithm is usually used for the final sound source distance determination [22]. Using Mel Frequency Cepstral Coefficients (MFCC), pretreatment and dynamic time warping (DTW) for speech detection, at three different SNR environment, were tested in typical real-time computer settings. The accuracy of the results reached 95%, the specificity reached 97%, and the sensitivity reached 76% [23]. Recurrent Neural Networks (RNN) can see all the information from the previous moment through cyclic feedback connections, and are suitable for modeling time series signals. |
| 2016-2017| "Acoustic Model" "Deep Neural Network" "Long Short-term Memory LSTM" "Automatic Speech Recognition Technology" "Sound Recognition" "Machine Learning" | Machine learning technology has long been the foundation of speech processing. Bayesian classification, decision tree, non-diffusion clustering, maximum class, maximum class, etc., were once part of the speech recognition system. |
With the rapid development of speech technology, models such as deep neural networks, convolutional neural networks, and recurrent neural networks have emerged, and new breakthroughs have been made in the training results of speech technology based on machine learning. Acoustic models based on deep neural networks have been successful in many speech recognition tasks. There are two common configurations of acoustic models based on deep neural networks: hybrid and tandem configurations. In the hybrid configuration, the neural network is used to predict the state of the context-dependent hidden Markov model, and the series configuration uses the deep neural network to perform nonlinear discriminant feature transformation [25-31]. Since the simple recurrent neural network (RNN) will have the problem of gradient disappearance, an improved model has emerged—a recursive structure based on long and short-term memory unit (LSTM) [32]. The end-to-end (E2E) multilingual automatic speech recognition (ASR) system aims to recognize multilingual speech within a unified framework and has been widely used in the field of automatic speech recognition [33]. As voice technologies become more mature in daily life, they will be able to provide human life with convenience and improve the quality of human life by understanding contextual information.

Speech emotion recognition plays a vital role in human-computer interaction. The current speech system uses bidirectional long- and short-term directed self-attention memory (BLSTM-DSA). Two-way long and short-term memory can make the structure more robust through the direction mechanism, and the autocorrelation of speech frames can be used to deal with the lack of information, so the self-attention memory mechanism is introduced. BLSTM-DSA performs well in speech emotion recognition tasks, especially in the recognition of happy and angry emotions, with the highest recognition accuracy [34].

In the past ten years, deep learning technology has continued to develop, and speech technology has gradually entered a new stage of rapid development. The evolution from deep neural networks to recurrent neural networks has made the historical memory of the machine longer. The emergence of the LSTM model better solves the problem of gradient disappearance in recurrent neural networks. With
the improvement of the algorithm model and the improvement of people's living standards, the "cochlear implant" technology has begun to be widely used in clinical practice, and has played a significant role in the language rehabilitation of the elderly and deaf-mute patients. In 2015, the rise of "end-to-end" technology laid the foundation for the further improvement of speech technology performance. In 2018, Google developed a more mature "end-to-end speech recognition model" and widely used it in automatic recognition systems, providing help for the optimization and improvement of intelligent speech technology. Faced with the more diverse needs of human beings, speech technology is innovating towards a more humane and intelligent direction, and significant improvements have been made in some emotion recognition scenarios.

6. Shortcomings and improvements

6.1. Insufficiency of speech technology development

(1) Inaccurate recognition caused by noise interference

In various speech recognition application scenarios, noise is an unavoidable interference factor. When using a speech navigation assistant on a highway, you will be exposed to noise caused by road friction, wind noise, and voices of passengers in the car. In a medical scenario, there will be noise from equipment and other patients in the same ward. These situations will cause a certain degree of interference to the accuracy of speech recognition.

For noise interference problems, the current mainstream processing methods are noise reduction technology based on deep neural networks and optimization methods based on acoustic features. The deep neural network itself has strong noise resistance, so this technology has achieved better results in high-noise environments. The optimization method of the acoustic model aims to use the invariance of convolution to overcome the diversity of the speech signal itself, and to adjust the multi-layer convolution unit structure of the model [35]. When the training set has a lot of data, the number of neurons in the hidden layer can be modified to prevent the network from overfitting, enhance the generalization ability and robustness of the neural network, and achieve the effect of noise reduction.

(2) Poor generalization ability and widespread ambiguity

There are non-standard expressions in Chinese such as polyphonic characters, puns, inverted sentences, dialects, and colloquialism, which may produce incorrect recognition results when the speech recognition system recognizes it. For example, people in Sichuan and Chongqing like to speak of "shoes" as "hai zi (A homonym for children)". When the speech system receives this vocabulary, it cannot correctly judge the meaning of the user.

At this stage, the most widely used method is to transform the features of the input content based on the deep learning unsupervised method, so as to improve the accuracy of feature classification in the recognition process. That is to say, pre-training on large-scale unsupervised corpus reduces the dependence on data in the process of corpus knowledge transfer, which can effectively speed up model convergence and solve the problems of poor generalization ability and ambiguity in the speech recognition process.

(3) Poor sound reception

In terms of speech collection, the effect of long-distance collection is not good, and there are interference factors such as reverberation and echo. When the user is cooking in the kitchen, he wants to use the voice assistant in the living room to learn how to make dishes. However, due to the distance problem, sometimes the voice assistant may not be able to recognize the user’s voice, or the recognized sentence may be incomplete and unable to do so. The user provides the right help.

For the problems of uncertain sound source, unsatisfactory sound reception, echo and noise in far-field recognition, it is not feasible to use traditional recognition methods. In recent years, with the development of far-field sound pickup technology, microphone array arrangements and software algorithms have become more and more abundant, and far-field sound pickup capabilities have improved significantly [36]. At present, the most widely used is the combination of multi-channel synchronous acquisition hardware R&D technology and microphone array arrangement algorithm.
6.2. Suggestions for the development of speech technology

(1) Suggestions on noise interference

In the current speech recognition system, there is a lack of an acoustic model optimization method that combines the invariance of convolution with a deep neural network. For the inevitable noise problem in the environment, a large amount of simulation training can be used to memorize the characteristics of the environmental noise and apply the training data to the speech system. For the sound interference of other non-primary speakers in the space, it can be judged by identifying the decibel level of the sound. Prioritize the recognition of the sound with the highest decibel, and ignore the weaker sound in the received sound.

(2) Suggestions on generalization ability and ambiguity

The current speech recognition system lacks a method that combines deep learning with pre-training models, which limits the new breakthroughs in accuracy of speech recognition technology to a certain extent. When conducting model training, look for a large number of testers with different speech tones and habits for radio training. For the recognition of dialects in different regions, more representative dialects are included in the training scope. The training vocabulary is not limited to simple daily conversations, and focuses on vocabulary that is different from Mandarin.

(3) Suggestions on the radio effect

Lack of a model that combines speech signals collected by multiple microphones with deep learning algorithms, and the recognition effect has a certain degree of room for improvement. Add multiple microphones for collecting sounds in the recognition system, and make the array of microphones face multiple directions to collect the speaker's speech from multiple sources. Combined with the deep learning algorithm model, the system can memorize the speaker's volume and sound source characteristics in different directions and different distances, so as to improve the completeness and accuracy of the remote radio.

7. Conclusion

In recent years, intelligent speech technology has gradually developed into one of the main development trends in the field of artificial intelligence, providing convenience for human production and life. Google proposed an end-to-end speech recognition model on November 15, 2018 [37], which triggered an upsurge in end-to-end speech recognition technology. Under the current development background, it is necessary to apply language adaptability training to the end-to-end multi-language automatic recognition system to improve its language recognition ability [38]. At the same time, in the development process, the deep neural network (DNN)-based modeling method is combined with the end-to-end speech recognition technology based on deep learning to improve the performance of intelligent speech technology to a certain extent.

At this time, the development of speech technology in intelligent voice assistants, speech synthesis technology, etc. is relatively high and has a wide range of applications. Most of the existing speech technologies rely on developers to input data for training to achieve the use effect, and cannot recognize the emotional color in human language. The Chinese culture is broad and profound, and the same sentences may not necessarily express the same purpose under different emotions. Therefore, the emotion in speech can better express the information hidden behind the text [39]. As the forefront of speech technology, neural language planning (NLP) aims to enable computers and systems to understand the true meaning of speech. With the development of artificial intelligence technology, machines are expected to obtain large-scale training data sets, which can automatically match related personalized services by reading the state of human speech, emotions, etc. And when users want to obtain a certain piece of information, they can be associated with each other. Provide users with multiple pieces of information.

Intelligent speech technology has played a huge role in scenarios such as home, office, and car, making people's lives more and more convenient, and sometimes it can even help people save time and complete tasks more efficiently. In the future, with the continuous exploration and improvement of
speech technology, more functions will be presented to users, which will better replace mechanized work and reduce some unnecessary troubles in daily life.

Reference

[1] Zhang Xiaolan, Wang Baolun. Research on the application of intelligent speech technology in teaching [J]. Modern Educational Technology, 2011, 21(11): 91-94+90.
[2] FERGUSON J D. Application of hidden Markov models to text and speech [EB]. 1980.
[3] Qin Fengzhi. Interpretation of the application of artificial intelligence speech transcription technology in conferences [J]. Electronic World, 2020(22): 190-191.
[4] Zhang Shuang. Development status, problems and solutions of speech recognition technology[J]. Journal of Physics: Conference Series,2020,1693(1).
[5] Dai Lirong, Zhang Shiliang, Huang Zhiying. Current status and prospects of speech recognition technology based on deep learning [J]. Data Collection and Processing, 2017, 32(02): 221-231.Jie Jiang, Harry Haoxiang Wang. Application intelligent search and recommendation system based on speech recognition technology[J]. International Journal of Speech Technology,2020 (prepublish).
[6] YU D, DENG L. Deep learning and its applications to signal and information processing[J]. IEEE
[7] Atieh Poushneh. Humanizing voice assistant: The impact of voice assistant personality on consumers’ attitudes and behaviors[J]. Journal of Retailing and Consumer Services,2021,58.
[8] Health and Medicine - Medical Technology; Data on Medical Technology Reported by Researchers at University of Sherbrooke (Speech Recognition for Medical Dictation: Overview in Quebec and Systematic Review) [J]. Journal of Engineering,2018.
[9] CHEN C M. CiteSpaceII: Detecting and visualizing emerging trends and transient patterns in scientific literature [J]. Journal of the American Society for Information Science and Technology, 2006 (3) : 359.
[10] Li Sun, Fan Zhiyan. Research on the multi-modal development trend of the intelligent voice industry under the AI+ trend [J]. Information and Communication Technology and Policy, 2019 (06): 17-21.
[11] Wang Haikun, Pan Jia, Liu Cong. Research progress and prospects of speech recognition technology[J]. Telecommunications Science, 2018, 34(02): 1-11.
[12] Chen Hangding. The team of the Key Laboratory of Language Acoustics and Content Understanding of the Chinese Academy of Sciences won the first place in the DCASE2019 Audio Scene Competition [J]. Network New Media Technology, 2019, 8(04): 66.
[13] Yu Tiecheng, Zhou Jianlai. The ears and mouths of computers-on computer speech recognition and synthesis technology [J]. China Information Herald, 1998(08): 39.
[14] Lin Deming, Chen Chaomei, Liu Zeyuan. Research on Zipf-Pareto Distribution of Co-cited Network Mediation[J]. Journal of Information, 2011, 30(1): 76-82.
[15] Ge Qian. Scientometric sorting by importance for literatures on life cycle assessments and some related methodological discussions[J]. The International Journal of Life Cycle Assessment,2014,19(7).
[16] Zhong Daidi. The perfect combination of computer and voice-CTI technology [J]. Journal of Beijing Broadcasting Institute (Natural Science Edition), 2002(03): 33-38.
[17] Wu Yale, Liu Xirui. Knowledge Graph Analysis of Domestic Speech Science Research——Based on the Comprehensive Application of CiteSpace and VOSviewer[J]. Journal of Zhengzhou University of Light Industry (Social Science Edition), 2020, 21(05): 93-102.
[18] Zhang Shuang. Development status, problems and solutions of speech recognition technology[J]. Journal of Physics: Conference Series,2020,1693(1).
[19] Wang Xingbao, Lei Qinhui, Mei Linhai, Zhang Ya, Xing Meng. Overview of the development trend of automobile voice interaction technology [J]. Automobile Digest, 2021(02): 9-15.
[20] Duke Mila Morais, Wolfe Jace, Schafer Erin. Recognition of Speech from the Television with Use of a Wireless Technology Designed for Cochlear Implants. [J]. Journal of the American
[21] Shulai Yin, Hui Fang, Xuehua Hou. Athlete’s respiratory frequency and physical energy consumption model based on speech recognition technology[J]. International Journal of Speech Technology, 2020 (prepubhlish).

[22] Yağanoğlu Mete. Real time wearable speech recognition system for deaf persons[J]. Computers and Electrical Engineering, 2021, 91.

[23] Pascale Fung, Dan Roth. Guest Editors Introduction: Machine Learning in Speech and Language Technologies[J]. Machine Learning, 2005, 60(1-3).

[24] G. E. Dahl, D. Y u, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” IEEE Trans. Audio, Speech, Language Process., vol. 20, no. 1, pp. 30–42, 2012.

[25] L. Deng, J. Li, J.-T. Huang, K. Y ao, D. Y u, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero, “Recent advances in deep learning for speech research at Microsoft,” in Proc. Int. Conf. Acoust., Speech, Signal Process., 2013, pp. 8604–8608.

[26] G. Hinton, L. Deng, D. Y u, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. V anhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition,” IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, 2012.

[27] T. Yoshioka, A. Ragni and M. J. F. Gales, "Investigation of unsupervised adaptation of DNN acoustic models with filter bank input," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 2014, pp. 6344-6348, doi: 10.1109/ICASSP.2014.6854825.

[28] A. Mohamed, G. E. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” IEEE Trans. Audio, Speech, Language Process., vol. 20, no. 1, pp. 14–22, 2012.

[29] H. Hermansky, D. Ellis, and S. Sharma, “Tandem connectionist feature extraction for conventional HMM systems,” in Proc. Int. Conf. Acoust., Speech, Signal Process., 2000, pp. 1635–1638.

[30] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, “Probabilistic and bottle-neck features for LVCSR of meetings,” in Proc. Int. Conf. Acoust., Speech, Signal Process., 2007, pp. IV–757–IV–760.

[31] Dai Lirong, Zhang Shiliang, Huang Zhiying. Current status and prospects of speech recognition technology based on deep learning [J]. Data Collection and Processing, 2017, 32(02): 221-231.

[32] Danyang LIU, Ji XU, Pengyuan ZHANG. End-to-End Multilingual Speech Recognition System with Language Supervision Training[J]. IEICE Transactions on Information and Systems, 2020, E103.D (6).

[33] Li Dongdong, Liu Jinlin, Yang Zhuo, Sun Linyu, Wang Zhe. Speech emotion recognition using recurrent neural networks with directional self-attention[J]. Expert Systems With Applications, 2021, 173.

[34] Zhang Shuang. Development status, problems and solutions of speech recognition technology[J]. Journal of Physics: Conference Series, 2020, 1693(1).

[35] Peng Hongsong, Li Hongbin, Li Li, Shui Guomei. Research on far-field speech recognition technology in artificial intelligence[J]. Digital Communication World, 2020(05): 193.

[36] He Y, Sainath T N, Prabhavalkar R, et al. Streaming End-to-end Speech Recognition for Mobile Devices[C]// ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

[37] Danyang LIU, Ji XU, Pengyuan ZHANG. End-to-End Multilingual Speech Recognition System with Language Supervision Training[J]. IEICE Transactions on Information and Systems, 2020, E103.D(6).

[38] Liu, M., Chen, H., Li, Y., & Zhang, F. (2015). Emotional tone-based audio continuous emotion recognition. In International conference on multimedia modeling (pp. 470–480). Springer.