Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/36/20230429

Human-computer interaction based on speech recognition

Ruixin Lu[1], Renjie Wei[2,4] and Jian Zhang[3]

[1]School of Artificial Intelligence, Jianghan University, WuHan, China
[2]School of Art and Design, Shanghai University of Engineering Science, ShangHai China
[3]School of software engineering, Henan University of Science and Technology, Luoyang, China

[4]mwellsu50098@student.napavalley.edu

Abstract. With the rapid change of The Times, language is no longer limited only in books, but gradually dedicates itself to reality. Speech recognition, in recent years, the cause of artificial intelligence and human-computer interaction continues to develop, all levels of life have its footprint, speech human-computer interaction leaf slowly began to integrate into the mainstream team of artificial intelligence. In general, speech recognition technology brings more convenience and naturalness to human-computer interaction, The benefits of human-computer interaction are not only reflected in improving the performance and efficiency of technical systems, but also in improving the user experience, fostering innovation, and promoting inclusive and sustainable development of society, and it has a positive impact in many fields, and with the continuous progress of technology, its application prospects will be broader. This paper makes a simple analysis and introduction of the role of speech recognition in human-computer interaction, expounds the key technologies, main algorithms and working principles of speech human-computer interaction, And some applications of human-computer interaction based on speech recognition in life and production fields. At the same time, the hidden problems and solutions are discussed, and includes the prospect of future speech human-computer interaction. Hope that this paper can bring some inspiration to relevant scientific research teams, assisting them to broaden a bright future.

Keywords: Speech Recognition, Artificial Intelligence, Human-Computer Interaction.

1. Introduction

With the development of artificial intelligence technologies such as deep learning and neural networks, the accuracy rate of speech recognition systems has been significantly improved. Advanced algorithms and models can better understand and interpret speech signals, thus improving the accuracy of speech recognition. Modern speech recognition technology can complete speech conversion in a short period of time and can respond quickly to users' commands and requests. Here's a look at the current situation and momentum. Improving accuracy: With the development of artificial intelligence technologies such as deep learning and neural networks, the accuracy of speech recognition systems has been significantly improved. Advanced algorithms and models can better understand and interpret speech signals, thus improving the accuracy of speech recognition [1].

DOI: 10.54254/2755-2721/36/20230429

Real-time and responsive: Real-time means that the speech recognition system can process the input speech signal in Real time (instant) or almost instant, and return the recognition result in a short time. In a real-time system, the response time should be short enough to produce a corresponding output immediately after the speech input ends. For many application scenarios, especially those that require immediate feedback (e.g., voice assistants, telephone interactions, real-time translation, etc.), real-time is critical.

Responsiveness: Responsiveness refers to the ability of the system to respond quickly to user input (voice). Even if the speech recognition system cannot process speech in real time, it should start processing immediately after receiving the voice input, and give the user certain feedback in a short time, such as progress tips or dynamic waveform charts [2]. Such feedback can let the user know that the system is processing their voice input and avoid giving the user a silent feeling of waiting [3]. Previous studies have mostly used microphone signal processing to isolate and analyze target speech, such as feature recognition and supervised machine learning through large amounts of training data. This method is suitable for the suppression of stationary noise, and it is difficult to meet the real-time processing requirements [4]. Next is the research content. "In this study, a real-time speech separation method based on the combination of optical camera and microphone array is invented. The method is divided into two steps. In the first step, computer vision technology is used together with the camera to detect and identify the object of interest, and determine the source Angle and distance. In the second step, microphone array beamforming technology is applied to enhance and separate the target speech. By using asynchronous updating function to combine beamforming control with speech processing, many problems of processing delay are avoided. This method has great prospects in machine language processing such as assisted listening systems or intelligent personal assistants [5].

Multilingual support: Speech recognition systems are increasingly supporting multiple languages. Many commercial speech recognition service providers have expanded their services to different languages and dialects around the world, enabling more people to use speech as a means of interaction. Both spoken language and symbols are made up of structured sub-lexical units. In speech signals, phonemes expand over time, while in symbols, visual sub lexical units such as position and hand shape are produced simultaneously.

Enhanced Context understanding: Speech recognition systems can not only translate speech into text, but also better understand context and semantics. This allows speech recognition to better handle complex voice commands and conversations and provide more accurate results. The extreme acoustic variability of speech is well known, which makes the proficiency of human speech perception all the more impressive [6]. "Speech perception, like any morphological perception, is contextual, which provides a way to normalize acoustic variability in speech signals. Acoustic environmental effects in speech perception have been extensively documented, but there is a lack of clear understanding of how these effects relate to each other across stimuli, timescales, and acoustic domains.

Expanding fields of application: Speech recognition technology is being used in an increasingly wide range of fields. In addition to traditional speech recognition applications such as voice assistant and voice transliteration, speech recognition is also used in smart home, car navigation, healthcare, customer service robots and other fields.

Cross-platform and mobile: Voice recognition technology has been widely used in mobile devices and smart phones, allowing users to interact with and control devices through voice. In addition, cross-platform speech recognition solutions are also evolving, enabling speech recognition to be applied to a variety of devices and systems.

Overall, speech recognition technology is constantly evolving and improving to provide a more accurate, real-time and diverse voice interaction experience. With the further development of artificial intelligence and machine learning, we can expect continuous breakthroughs and innovations in speech recognition technology in the future [7]. And research on speech recognition will make efforts to explore related fields: First, researchers can evaluate and compare various speech recognition systems and analyze their performance differences in different tasks or scenarios. Such work can help the research community understand the current state of technological development and provide guidance for system

DOI: 10.54254/2755-2721/36/20230429

selection in practical applications. Second, researchers can apply speech recognition systems to specific fields, such as healthcare, intelligent assistants, voice control, etc., and explore practical application effects and challenges in these fields [8]. Such research can promote the application of speech recognition technology in practical scenarios and provide support for the development of related fields. Finally, researchers can improve the existing speech recognition systems, such as proposing new feature extraction methods, acoustic model optimization strategies, and joint training methods for acoustic and language models.

2. Human-computer interaction based on speech recognition

2.1. Key technologies

By analyzing the input end of the man-machine interaction interface of energy management system (EMS) and comparing the characteristics of several input tools, the introduction of speech recognition technology is proposed. In the improved man-machine interface realizes a new assignment of input task to the three types of voice, mouse and keyboard input to achieve simplified task coefficient difficulty forms an input mode based on the cooperation of voice and mouse in command and control [9, 10]. Compared with the traditional man-machine interface the improved man-machine interface has obvious advantages in operation efficiency, freedom of input and display ability of output and its advantages are illustrated by experiments.

The following is a comparison of the command and control input tool technologies

Table 1. Comparison of the input tools in the command and control mode.

Output tool	property			
	Cognitive load	Operational efficiency	naturality	Application range
mouse	low	low	bad	overall
Keyboard (shortcut)	lower	Extremely high	bad	bad
voice	higher	height	good	common

Then there is the comparison of voice input tool technologies.

Table 2. Comparison of the input tools in the text input mode.

Output tool	property		
	Cognitive load	Operational efficiency	naturality
Mouse (soft keyboard)	lower	Extremely low	bad
Keyboard	lower	high	bad
voice	low	low	good

The key technologies of speech recognition include:

1) Front-end processing: Pre-processing the original speech signal, including sampling rate conversion, noise reduction, speech endpoint detection, etc., to extract clean and effective speech features. 2) Feature extraction: the speech signal is converted into feature sequences, commonly used features include Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC) and so on. 3) Modeling method: Use Hidden Markov Model (HMM), Deep Neural Networks (DNN) and other methods to model the acoustic characteristics and language model of speech signals. 4) Language model: According to different application scenarios, design and train language models to enhance the speech recognition system's understanding of semantics and context. 5) Pronunciation dictionary: contains the pronunciation information of each word, which is used to map the speech signal to the corresponding

DOI: 10.54254/2755-2721/36/20230429

word. 6) Morphological processing: consider the deformation rules of words in order to better adapt to the variations of daily spoken language, such as sound change, linking and so on. 7) End-to-end model: By directly mapping the input speech signal to the output text, the acoustic model, pronunciation model and language model that need to be separated in the traditional system are avoided [11]. These key technologies are often used in combination with system optimization and adjustments to improve the accuracy and performance of speech recognition systems.

2.2. Main algorithms

Speech recognition uses a variety of algorithms and techniques in human-computer interaction, some of the main algorithms are listed below: The main algorithms for speech recognition include: 1) Hidden Markov Model (HMM): HMM is a commonly used statistical model to represent the probability distribution of acoustic and language models. It assumes that speech signals are generated by a series of hidden states, and that the transitions between states conform to Markov processes [12]. 2) Recurrent Neural Networks (RNNS): RNNS are a special class of neural networks that can process data with temporal relationships. In speech recognition, RNNS are often used to model the timing relationships of speech signals, such as passing information from a previous speech segment to a later segment. 3) Convolutional Neural Networks (CNNS) : CNNS are widely used in image processing, but can also be used for speech recognition tasks. Through convolutional operations, CNNS can extract features in the speech spectrum, such as phonemes, speech fragments, or speech paragraphs. 4) Deep Neural Networks (DNN): DNN is a multi-layer neural network structure that models complex features of speech signals through multiple hidden layers. DNNS are often used in acoustic modeling in speech recognition to extract high-level abstract features from the speech spectrum [13].

In addition to the above algorithms, there are many other techniques and methods that are used in speech recognition. These algorithms and models can be used in combination to improve the performance and accuracy of speech recognition systems.

3. Workflow

Speech recognition technology basic principle and process introduction speech recognition system function division by speech signal preprocessing, feature extraction, pattern matching composed of three parts. The first step of preprocessing, mainly A/D transformation, pre-emphasis and endpoint detection part. After the pre-processing of the speech signal, to carry out the second step feature extraction, the process is to extract the required feature parameters in the original speech signal, so as to obtain the feature vector sequence, after the feature extraction is completed, the next is the core of speech recognition, that is, the third step pattern matching, that is, pattern system recognition. The step frame diagram is as follows [14]. The specific flow of voice interaction was shown in the Figure 1

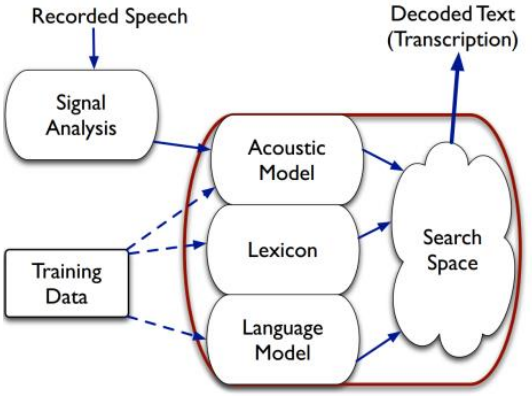

Figure 1. Speech recognition flow chart.

DOI: 10.54254/2755-2721/36/20230429

More detailed is divided into signal acquisition, front-end processing, feature extraction and acoustic model training and reasoning [15]. 1) Signal acquisition: First, the speech signal is collected through the microphone or other recording equipment. These signals can be the sounds of a single person speaking or multiple people interacting. 2) Front-end processing: In this step, the collected speech signal is preprocessed to remove noise, reduce interference factors such as echoes, and adjust the audio quality of the signal to make it suitable for subsequent processing. 3) Feature extraction: Next, features are extracted from the speech signal that has been processed by the front end. Technology such as short-time Fourier Transform (STFT) is usually used to convert the speech signal into a time-frequency representation, and then extract the features of the sound based on this time-frequency information, such as the Maier frequency cestrum coefficient (MFCC). 4) Acoustic model training and reasoning: Finally, using previously labeled speech data, the acoustic model is trained based on machine learning algorithms. Common models include Hidden Markov models (HMM) and deep learning models (such as recurrent neural networks or convolutional neural networks).

4. Applications of human-computer interaction based on speech recognition

4.1. Application of speech recognition in life

At present, voice interaction has long become the most widely used way of interaction in people's lives. Voice interaction on the hardware requirements are not high, the use of speakers and microphones can be achieved, no need to operate the controller, just through the dialogue can direct the machine to achieve the given instructions. With voice interaction, we can talk directly to the device and command it to perform a specific task. This directness and convenience make voice interaction widely used in our daily life.

Intelligent assistant, a smart device that integrates technologies such as speech recognition, natural language processing and machine learning, is gradually entering People's Daily life and bringing unprecedented convenience. They can chat with users, consult, query, but also self-learning and training, after several years of continuous development of artificial intelligence technology, the current intelligent voice recognition assistant can easily achieve voice interaction with users and communication reply, through speech and semantic recognition, voice recognition assistant constantly understand users, calculate user needs [16]. For example, Xiao Ai is able to answer questions, provide information and perform tasks through voice commands. Users can communicate with the assistant through voice interaction, such as checking the weather, playing music, sending messages and so on.

It can use voice recognition technology to convert the user's voice output into text, and then use natural language technology to process it, analyze the user's intention, and then complete subsequent related tasks.

In the past 20 years, smart home only looks advanced and convenient, in fact, many simple problems are complicated, and now, the "help" of intelligent algorithms, the general model support of big data statistics, the personalized customization of blockchain and its high consensus, and the new interactive experience brought by pattern recognition, artificial intelligence has brought new breakthroughs for the development of smart home. To solve the "dilemma" of smart home [16]. Application of voice interaction in intelligent furniture was shown in the Figure 2.

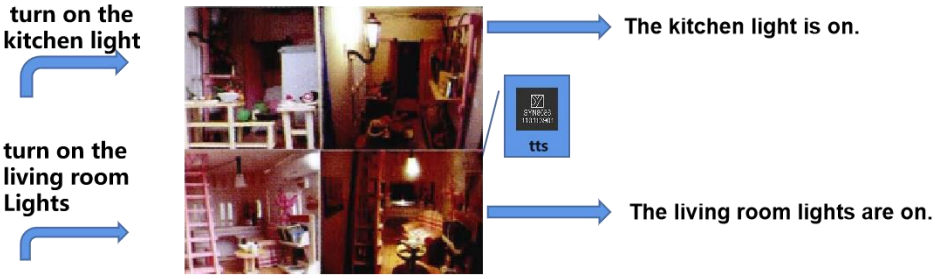

Figure 2. Live demo picture.

DOI: 10.54254/2755-2721/36/20230429

Smart home system, not only computer and communication technology, but also artificial intelligence technology. The core lies in the software of artificial intelligence chip and artificial intelligence algorithm. Smart home from the technical realization can be seen as "an application of the Internet of Things technology, and integration and artificial intelligence technology.

At present, smart home in the application of artificial intelligence technology can mainly achieve several interaction mode methods, such as the application of artificial intelligence technology can mainly achieve several interaction modes, such as touch interaction, voice interaction, body sensing interaction, AR (augmented reality) interaction/VR (virtual reality) interaction and so on. Augmented reality (AR) interaction/VR (virtual reality) interaction, and even some high-tech brain-computer interaction.

Among these interaction modes, voice interaction is more natural and concise in practical applications, and it can be used in a variety of ways: The advantages of more efficient input have been widely used in smart homes. Before the function is put into use, it must have full intelligent speech recognition, acoustic processing, semantic understanding and speech synthesis and other functions to ensure that every detail is carefully processed to improve the practicality and efficiency of smart home. Through voice interaction technology, smart home becomes more convenient, intelligent and humanized. Residents can control and manage various devices and systems in their homes through simple voice commands, improving the quality of life and living experience. With the continuous development of technology, we can expect more voice interactive applications for smart home devices and services.

4.2. Applications in gaming

With the development of artificial intelligence, games have also begun to develop into intelligent games. For example, many nurturing games rely on intelligent carriers to complete the game. The intelligent carriers in the game have thoughts, emotions, will feel hungry and laugh happily [17]. Voice interaction technology in digital games includes the use of voice for calculation processing, or for players' input and output to video games, but does not include voice communication between players [18]. The application of voice interaction in games is gradually increasing, providing players with a more immersive and interactive game experience.

Especially in cooperative multiplayer games, voice interaction plays an important role in cooperative multiplayer games. Players can use voice to communicate and collaborate with teammates in real time, formulate strategies, assign tasks, and defeat enemies together. This real-time voice communication can greatly improve team cooperation and combat efficiency in the game. The game provides a voice command function, through which players can control the characters in the game or perform specific game actions. For example, the player can use voice commands to get the character to attack, jump, use special skills, etc. In Yasuhati, the player's volume controls the character's forward walk: the louder the voice, the faster the walk. Above a certain range, the character starts to jump [19]. The control of voice input is natural and straightforward, and because it is easy to master. While it's easy to gibberish when playing a game using voice interaction, it can lead to an unexpectedly interesting gameplay experience [20]. Voice interaction can make a game more interactive and customizable. Players can customize the game experience by using voice interaction to customize their character's appearance, attributes, or make game Settings.

Overall, voice interaction provides a more intuitive, convenient and immersive experience for players in the game. Through voice interaction, players are better able to interact with the game world and other players, increasing the fun and interactivity of the game. As technology advances, we can expect more innovative applications of voice interaction in games. he application of voice interaction in games was shown in the Figure 3

DOI: 10.54254/2755-2721/36/20230429

Figure 3. Voice interaction and game scene diagram.

5. Discussion

This paper mainly discusses some key technologies of speech interaction and specific applications in some fields. Speech recognition technology has experienced the development from pattern matching to HMM, then to deep learning and the proposal of end-to-end model. The evolution of these technologies has improved the accuracy of speech recognition. In the future, with the continuous innovation of technology and the abundance of data resources, speech recognition technology will continue to move towards higher accuracy. Speech recognition has been widely used in some fields, including voice assistant and identity recognition in daily life. The application of voice interaction has brought great convenience to people's life and production, and improved production efficiency.

However, at present, speech recognition is also faced with some problems, one is the sound quality and environmental interference speech recognition system is highly sensitive to sound quality and environmental interference. For example, noise, echo and other factors will cause the quality of the speech signal to decline, thus affecting the accuracy of speech recognition. The other is speech variation and intonation difference: there are large differences in speech, pronunciation and intonation between different individuals, which will negatively affect the accuracy of speech recognition. The third is contextual semantic understanding: speech recognition systems need not only to recognize and transcribe speech signals, but also to understand and interpret their meaning and context. This includes semantic parsing and contextual understanding of speech signals in order to more accurately understand and generate speech content.

In order to solve these problems, using deep learning techniques, training and augmenting diverse data is essential. Deep learning has already achieved great success in speech recognition. In particular, using recurrent neural networks (RNNS) or variations thereof, such as Long Short-term memory networks (LSTMS) and gated cyclic units (GRUs), speech signals can be modeled and their long-term dependencies captured. Convolutional neural networks (CNNS) can also be used for speech feature extraction and preprocessing. These methods can greatly improve the accuracy of speech recognition, so as to achieve some breakthroughs in problems.

After years of development, speech recognition technology has made important breakthroughs and achievements. However, with the rapid development of artificial intelligence and machine learning, there is still great potential and room for future speech recognition technology. We can make the following prospects for speech recognition.

5.1. Further development of deep learning

Deep learning, as an important technology in the field of speech recognition, will continue to promote the development of speech recognition. In the future, more powerful and efficient deep learning networks will be developed to improve the accuracy of speech recognition.

5.2. The application of cross-modal fusion

Speech recognition can be combined with other modes of perception, such as images, gestures and biometrics. Future speech recognition systems will pay more attention to the integration of multi-modal information to provide more accurate, comprehensive and diversified speech understanding and recognition capabilities.

DOI: 10.54254/2755-2721/36/20230429

5.3. In-depth speech content understanding

Future speech recognition systems will be able to understand and interpret speech content in greater depth, such as more accurate and comprehensive analysis of emotion, intent, and context. The speech recognition technology of the future will continue to evolve and innovate to provide people with a more powerful, intelligent and natural voice interaction experience. This will promote the wide application of speech recognition in various fields and bring more convenience and efficiency to people's life and work. We look forward to further development of speech recognition technology in the future to make it an integral part of our daily lives.

6. Conclusion

This study provides some novel insights into the literature on speech interaction, which has been understudied in previous studies. The first contribution of this study is to conceptualize speech interaction by integrating speech interaction theory. Several theoretical implications are provided in this paper. First, our study is the first to explain that speech recognition technology has gone through the transition from pattern matching to HMM. Second, although speech interaction has been extensively studied in speech separation and stationary noise, this study not only demonstrates the study of non-stationary noise, but also demonstrates that speech signals can be modeled using recurrent neural networks (RNNS) or variations of them, such as Long short-term memory networks (LSTM) and gated cyclic units (GRUs). Finally, this paper provides a detailed explanation of the application of speech interaction in real life. There is no doubt that speech recognition plays a pivotal role in human-computer interaction. Whether it is daily life, scientific and technological development, or deep learning, speech recognition enables many complex problems to be solved easily, and opens up a broader world for human beings. Voice is everywhere, and speech recognition enables people to communicate and control with computers or other intelligent devices through oral language, which can greatly facilitate life and promote the development of science and technology.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Villameriel S, et al. Language modality shapes the dynamics of word and sign recognition. Cognition. 2019; 191:103979.
[2] Stilp C. Acoustic context effects in speech perception. Wiley Interdiscip Rev Cogn Sci. 2020; 11(1):e1517.
[3] Cowan T, et al. Masked-Speech recognition for linguistically diverse populations: A Focused Review and Suggestions for the Future. J Speech Lang Hear Res. 2022; 65(8):3195-3216.
[4] Vermiglio AJ, et al. Diagnostic Accuracy of the AzBio Speech Recognition in Noise Test. J Speech Lang Hear Res. 2021; 64(8):3303-3316.
[5] Xu S, et al. Research on Human-computer intelligent interaction based on speech recognition and natural language processing conversation flow. Machinery & Electronics 39.07(2021):65-69.
[6] Lu Z, et al." Human-computer interactive speech recognition development and analysis of military applications." Ordnance Industrial Automation. 2023, 42(04):21-25.
[7] Cheng H., et al. Implementation of speech recognition technology based on Linux platform. Internet of Things Technology. 2022, 12(10):89-91.
[8] Tao J., et al. Human-Computer interaction oriented to Virtual-real fusion. Journal of Image and Graphics. 2023, 28(06):1513-1542.
[9] Zhang H., et al. Research and implementation of intelligent dialogue system based on speech recognition." Journal of Shenyang Normal University (Natural Science Edition) 2022, 40(05):446-450.
[10] Yu K, et al.. Speech recognition and end-to-end technology status and prospects [J]. Application of Computer Systems, 2021, 30(3):14-23

DOI: 10.54254/2755-2721/36/20230429

[11] He Y., et al. A scene Recognition Method based on HMM [J]. Computer Science, 2011, 04:254-256.
[12] Liu Y, et al. Text information Extraction based on Hidden Markov Model [J]. Journal of System Simulation,2004(03):507-51
[13] Yu X. Development and application of speech recognition technology. Computer Times. 2019, 11:28-31.
[14] Xu S, et al." Research on human-computer intelligent interaction method based on speech recognition and natural language processing conversation Flow. Machinery & Electronics 2021, 39(07):65-69.
[15] Lu Z., et al Human-computer interactive speech recognition development and military application analysis. Ordnance Industrial Automation 2023, 42(04):21-25.
[16] Liu CF, et al. A real-time speech separation method based on camera and microphone array sensors fusion approach. Sensors (Basel). 2020, 20(12):3527.
[17] Zhu Q, et al. Speech recognition technology is applied in EMS man-machine interactive study [J]. Automation of electric power systems, 2008, 32 (13):45-48.
[18] Zhang H. Localization voice interaction technology based on the hardware in the application of the smart home system [J]. Journal of electronics science and technology, 2013(6):1.
[19] Li M, et al.. Intelligent household scenario the children's game of interaction design research [J]. Journal of packaging engineering, 2022,16:68-75.
[20] Wang W. Artificial intelligence to the subjective influence [J]. Journal of reform and opening, 2018(14):113-114.