Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/17/20230907

The application and challenges of artificial intelligence in speech recognition

Bohan Wang

Department of Computer Science and Engineering, University of Connecticut, Connecticut, 06269, USA

bohan.wang@uconn.edu

Abstract. This paper provides an overview of artificial intelligence (AI) and speech recognition technology, including its history, applications, challenges, and future prospects. AI-powered speech recognition technology has significantly improved over the years, and it is used in various applications, such as virtual assistants, voice-activated devices, and dictation software. The technology leverages machine learning algorithms that are trained on vast amounts of speech data to recognize and interpret human speech with accuracy levels that are comparable to those of humans. However, the technology still faces many challenges, such as speech variability and background noise, which make it challenging to develop speech recognition algorithms that can accurately recognize all types of speech. The article provides a comprehensive review of the technical aspects of automatic speech recognition, including the process involved, the algorithms used, and the challenges and opportunities for future research in this area. The paper also discusses the architecture of automatic speech recognition (ASR) systems and the main components that make up the system. The authors explain that ASR systems consist of three main components: the acoustic model, the language model, and the decoder. They also discuss the challenges that ASR systems face, such as speaker variability, noise, and limited vocabulary. Overall, this paper provides a detailed introduction to AI and speech recognition technology and its potential for various industries.

Keywords: speech recognition, ASR system, Mel-frequency cepstral coefficients (MFCCs), linear predictive coding (LPC), deep neural network (DNN).

1. Introduction

Artificial Intelligence (AI) has revolutionized many industries [1-4], and the field of speech recognition is no exception. AI-powered speech recognition technology has significantly improved over the years, and it has become an integral part of many applications such as virtual assistants, voice-activated devices, and dictation software. This paper aims providing an overview of AI and speech recognition, discussing the technology's history, applications, challenges, and future prospects.

Speech recognition technology dates back to the 1950s when Bell Laboratories developed the first speech recognition system. However, the technology was still in its infancy and was not practical for commercial use. Over the years, speech recognition technology has evolved, and the introduction of AI has led to significant advancements in the field.

DOI: 10.54254/2755-2721/17/20230907

AI-powered speech recognition technology leverages machine learning algorithms that are trained on vast amounts of speech data to recognize and interpret human speech. The technology works by breaking down audio signals into smaller components, such as phonemes, and then using statistical models to identify the words that the speaker is saying. The precision of speech recognition technology has undergone a significant improvement, enabling it to accurately comprehend human speech with accuracy levels that are comparable to those of humans. One of the most significant applications of AI and speech recognition technology is virtual assistants, such as Apple's Siri, Amazon's Alexa, and Google Assistant. These virtual assistants use speech recognition technology to understand user commands and respond with relevant information or actions. They can perform a wide range of tasks, such as setting reminders, making phone calls, and playing music. Another significant application of AI and speech recognition technology is in voice-activated devices. These devices include smart home devices, such as smart thermostats and smart speakers, and are designed to respond to voice commands. The application of speech recognition technology enables them to decipher and execute user commands, including but not limited to, turning off lights and playing music. Furthermore, the applicability of AI and speech recognition technology extends to the healthcare industry. The technology can be used to develop speech-based diagnostics tools that can detect speech-related disorders, such as dysarthria and stuttering. It can also be used to develop speech-to-text software that can transcribe medical dictation, allowing doctors to spend more time with patients and less time on paperwork.

However, despite the significant advancements in AI and speech recognition technology, the technology still faces many challenges. One of the most significant challenges is speech variability. Human speech is highly variable, and different people may pronounce words differently or use different dialects. This variability makes it challenging to develop speech recognition algorithms that can accurately recognize all types of speech. Another challenge is background noise. Speech recognition algorithms are designed to recognize speech in quiet environments, but in real-world situations, there may be significant background noise, making it challenging to interpret human speech accurately.

The article provides a comprehensive review of the technical aspects of automatic speech recognition, including its history, the algorithms used, and the challenges and opportunities for future research in this area.

2. Methodology

The automatic speech recognition (ASR) system is a complex process that involves several stages. First, the speech signal is converted into a digital format and then preprocessed to enhance the quality of the signal by removing noise and filtering out unwanted frequencies. The preprocessed signal is then transformed into a feature representation, such as Mel-frequency cepstral coefficients (MFCCs) or linear predictive coding (LPC), which captures the relevant characteristics of the speech signal [5]. The feature representation is then used as input to a machine learning model, such as a deep neural network (DNN), which is trained on a large dataset of labeled speech data to recognize speech patterns. During recognition, the DNN produces a sequence of probability distributions over a set of phonemes or words, which are used to construct the most likely transcription of the speech [5]. The decoding process involves searching for the most probable sequence of words or phonemes that correspond to the observed speech signal.

Speech recognition process has been separated into five different models, 1. Acoustic model, 2. Language model, 3. Trigram model, 4. Class model, and 5. Source channel model [6]. First of all, Acoustic model indicates the acoustic sound of a language, and it can also be trained for recognizing specific user’s speech pattern and the traits of the acoustic environment of the user. In addition, the lexical model provides a list of vast of vocabularies, and their pronounce. Language model provides multiple combinations of words and languages, in order to help the recognizer to define a word for the vocabulary lists and select it. A Trigram model examine the probability of speeding, which means the process of guessing what language or word will appear next depends on the history of previous

DOI: 10.54254/2755-2721/17/20230907

speeches. The fourth model is called class model, it serves the purpose of classifying words and creating sets of words since a single word could belongs to different classes, so the classing model makes deep morphological analysis upon words and its information. Last but not the least, Source channel model helps minimize the error rate that has been recognized and choose the sequence with max posterior distribution [7].

Saliha Benkerzaz et al. describe the architecture of ASR systems and the main components that make up the system. The authors explain that ASR systems consist of three main components: the acoustic model, the language model, and the decoder. The acoustic model represents the acoustic characteristics of speech signals and converts them into a sequence of phonemes. The language model represents the probability distribution of sequences of words in a language and determines the most probable word sequence given the acoustic input. The decoder combines the outputs of the acoustic and language models to produce the final recognition output [7]. The authors also discuss the challenges that ASR systems face, such as speaker variability, noise, and limited vocabulary, and how these challenges can be addressed through techniques such as feature extraction, acoustic modeling, language modeling, and adaptation [7].

3. Application and discussion

3.1. Amazon alexa

Amazon Alexa is an AI-powered voice assistant developed by Amazon, employing Natural Language Processing (NLP) and speech recognition technology to comprehend and address user queries [8]. It can perform a range of tasks, including answering questions, setting reminders, playing music, controlling smart home devices, and ordering products from Amazon. Alexa's speech recognition technology is based on deep learning algorithms, which enables it to recognize and respond to a wide range of accents and dialects. The success of it can be attributed to its advanced speech recognition capabilities. By using machine learning algorithms, Alexa can improve its comprehension of diverse accents and languages over time, making it popular among users from different parts of the world.

3.2. Google assistant

Google Assistant is an AI-powered voice assistant developed by Google. The function of it is similar to the Amazon Alexa mentioned above. Google Assistant is integrated into a range of devices, including smartphones, smart speakers, and smart displays. It can perform tasks such as sending messages, making phone calls, and providing weather and traffic updates. Google Assistant's speech recognition technology is one of the most advanced in the industry. It uses a combination of acoustic models, language models, and machine learning algorithms to recognize and understand user requests [9]. Moreover, Google has invested substantially in enhancing its NLP capabilities, making it increasingly conversational and human-like, thereby making it a popular choice among users who seek a more natural interaction with their voice assistants.

3.3. Dragon dictation

Dragon Dictation is an AI-powered speech recognition tool developed by Nuance Communications. It is designed to transcribe spoken words into text. Dragon Dictation employs a combination of acoustic models, language models, and machine learning algorithms to recognize and interpret spoken words. It can be used for a range of tasks, including writing emails, composing documents, and taking notes [10]. Dragon Dictation's speech recognition technology is highly accurate and efficient. It uses a deep learning algorithm that can adapt to different accents and dialects, making it popular among users from different parts of the world. The tool is also highly customizable, allowing users to create custom commands and shortcuts to improve their productivity. Overall, Dragon Dictation is a potent tool that caters to the needs of inviduals who require spoken word transcription into written text.

DOI: 10.54254/2755-2721/17/20230907

3.4. Discussion

Google Assistant, Amazon Alexa, and Dragon Dictation are three popular virtual assistant technologies that have transformed the way people interact with the devices. Google Assistant is known for its natural language processing capabilities, which allow for conversational interactions. It can assist with a range of tasks, including setting reminders, answering questions, playing music, and controlling smart home devices. Alexa also offers a range of functions, with integration into Amazon services such as Amazon Music and Amazon Prime and boasts a large number of third-party skills. Dragon Dictation, on the other hand, is a speech recognition software designed for dictation and transcription purposes, making it popular among professionals.

Given the continual advancements in AI and natural language processing technology, the future of virtual assistants appears promising. It can be anticipated these technologies becoming increasingly conversational and capable of performing more complex tasks. Furthermore, as more third-party services integrate with virtual assistants, their overall value to users will continue to increase. Personalization through the use of machine learning and data analysis may also lead to more intuitive interactions with individual users.

4. Conclusion

AI systems and speech recognition make it much more convenient for users to work with technology without requiring for a keyboard or mouse. These technologies can perform tasks quickly and accurately, improving productivity and reducing errors. AI systems and speech recognition can help people with disabilities to interact with technology more easily.AI systems and speech recognition have transformed the way people interact with technology. They have made it possible for users to communicate with devices using natural language, opening up a world of possibilities for personal and professional applications. The future of AI systems and speech recognition is expected to be bright. With advancements in machine learning and natural language processing, these technologies are likely to become more sophisticated and accurate. The integration of these technologies into various industries, such as healthcare and education, is also expected to increase, improving efficiency and productivity. However, it is important to address the cons associated with these technologies to ensure their safe and ethical use in the future.

References

[1] Kayalibay B Jensen G van der Smagt P 2017 CNN-based segmentation of medical imaging data arXiv preprint arXiv:1701.03056.
[2] Yu Q Wang J Jin Z et al. 2022 Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training Biomedical Signal Processing and Contro 72 103323.
[3] Hamarneh G Yang J McIntosh C et al 2005 3D live-wire-based semi-automatic segmentation of medical images Medical Imaging 2005: Image Processing SPIE 5747 pp 1597-1603.
[4] Govindan K 2022 How artificial intelligence drives sustainable frugal innovation: A multitheoretical perspective IEEE Transactions on Engineering Management.
[5] Arora S J and Rishi P S 2012 Automatic speech recognition: a review International Journal of Computer Applications 60.9.
[6] Choudhary A and Ravi K 2012 Process speech recognition system using artificial intelligence technique International Journal of Soft Computing and Engineering (IJSCE) 2.
[7] Benkerzaz S Youssef E and Abdeslam D 2019 A study on automatic speech recognition Journal of Information Technology Review 10.3 77-85.
[8] Kumar, Deepak et al 2018 Skill squatting attacks on Amazon Alexa 27th {USENIX} Security Symposium ({USENIX} Security 18).
[9] Velikovich L et al 2018 Semantic Lattice Processing in Contextual Automatic Speech Recognition for Google Assistant Interspeech.

DOI: 10.54254/2755-2721/17/20230907

[10] MacArthur C A and Albert R C 2004 Dictation and speech recognition technology as test accommodations Exceptional Children 71.1 43-58.