Proceedings of Urban Intelligence: Machine Learning in Smart City Solutions - CONFSEML 2024
DOI: 10.54254/2755-2721/65/20240493

Automatic speech recognition technology: History, applications and improvements

Qiyu Liang

School of Technology, Beijing Forestry University, Beijing, China, 100091

likeastar@bjfu.edu.cn

Abstract. In today's world, automatic speech recognition(ASR) has been an important part of artificial intelligence. It has been recognized as an extremely difficult highly challenging high-tech topic. It mainly converts the vocabulary content in human speech into computer-readable input, which is generally understandable text content, and may also be binary encoding or character sequences. Since the 1950s, ASR has been continuously developing from simple systems for pronunciation of 10 English numbers to the rise of multiple frameworks and different neural networks. The process of ASR is constantly becoming diversified and specialized. Based on the analysis of existing literature, this article will briefly describe the history of speech recognition technology, the current development status of speech recognition, various applications in daily life and advanced areas, and methods for improvements. It indicates that nowadays automatic technology has become an essential part in people's daily lives. Simple methods for eliminating echoes and noise to improve system performance and user experience are also an important part that should be considered in the use of ASR.

Keywords: Automatic speech recognition, Markov model, improvement

1. Introduction

Today, with the rise of computer technology, more and more people are beginning to invest the research of more interaction between computers and humans, and speech recognition technology has become a hot topic. In daily communication, humans perceive others' expressions through hearing, touching, and vision, such as listening to others' words, watching their body movements, and feeling their inner emotions through touching. On the other hand, computers are different. In the computer world, knowledge and expression are mostly generated by binary code, and how to convert human communication speech into binary code inside the computer has become an important topic. Speech recognition is a cutting-edge technology that integrates multidisciplinary knowledge, covering basic and cutting-edge disciplines such as mathematics and statistics, acoustics and linguistics, computer and artificial intelligence, and is a key link in human-computer natural interaction technology. However, speech recognition technology has not shown great potential in recent applications, and its various shortcomings still make its development relatively slow. After the invention and application of various neural network systems today, there has been relatively rapid development and application, but it still shows shortcomings in dealing with noise, interference, and other aspects. Based on the needs of the automation profession and interest in researching automatic speech recognition, after an extensive

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/65/20240493

literature review, this article will elaborate on the continuous development of automatic speech recognition (ASR) technology over the past 70 years, especially its inception in the 1950s. This paper will briefly describe the development history of ASR technology and the application prospect of ASR technology, provide improvement measures and ideas, enhance the understanding of ASR technology, and provide guidance for future researchers to enter into in-depth research on ASR.

2. Developments of ASR

ASR has undergone years of development, from the application of simple English words to today's intelligent speech systems, and its years of development are also one of the major topics.

2.1. The beginning of modern ASR

Bell Laboratories' Davis, Biddulph, and Balashek developed a system for isolated digit recognition for a single speaker in 1952. The system used formant frequencies that were measured or calculated during each digit's vowel regions [1]. Subsequently, in the United Kingdom, Denes and associates created the initial computer voice recognition system in 1960. After, extensive research on speech recognition was initiated after the 70s of the last century, and substantial progress has been made in the recognition of small vocabulary and isolated words. Since the 80s of the last century, the focus of speech recognition research has changed to continuous speech recognition little by little with large vocabulary and non-specific people. Simultaneously, speech recognition has additionally gone through tremendous changes in the research ideas, from the customary specialized thoughts in view of standard format matching to the specialized thoughts in light of factual models. Once more, a few specialists in the business set forward the specialized thought of bringing brain network innovation into discourse acknowledgment issues. Since the 90s of the last century, there has been no major breakthrough in the system framework of speech recognition. However, there has been great progress in the application and commercialization of speech recognition technology. For example, a program funded by the U.S. Defense Vision Research Projects Agency in the 70s to support research and development of language understanding systems. In the 90s, the program was still ongoing, and its exploration focus had moved to the regular language handling part of the acknowledgment gadget, and the recognition task was set as “air travel information retrieval”.

2.2. Application of HMM

The identification algorithm based on the implicit Markov model (HMM) is a voice recognition algorithm in the field of voice recognition in the 1980s. This algorithm establishes a statistical model of identification entries through data statistics on a large amount of voice data, and then extracts special Symptoms, match these models to obtain recognition results by comparing scores. Through a large number of voice, a stable statistical model can be obtained, which can adapt to various emergencies in actual voice. Therefore, the HMM algorithm has good identification and noise resistance. The recognition system based on HMM technology can be used for non -specialty and does not require users to train in advance. His drawback is that a larger voice library is required for the statistical model's establishment. There is a lot of real labor involved in this. Relatively significant amounts of storage and matching computations are needed for the model, including the feature vector's output probability calculation, and DSPs with a certain capacity SRAM are usually required to complete [2]. HMM technology has been widely used in the fields of voice recognition, machine translation, Chinese segmentation, naming entity recognition, word marking, and genetic recognition. Although in today's era, the rise of many new neural network technology has made its importance no longer the year, but it still has high learning value as a technology that has been used for many years.

2.3. Current situation of ASR

Today, ASR technology has shifted to the era of end-to-end. The end-to-end model is a system that directly maps the input audio sequence to the word sequence or other words. The components of the majority of end-to-end speech recognition models are as follows: The voice input sequence is mapped to

DOI: 10.54254/2755-2721/65/20240493

a feature sequence by the encoder; the language and feature sequence alignment is realized by the aligner; and the final identification result is decoded by the decoder. Please be aware that this separation is not always there since a designed modular system, when compared end-to-end, is a full structure and it is typically very difficult to determine which element performs which sub-task. Unlike the HMM-based model, which consists of many modules, the end-to-end model uses a deep network in place of several modules to realize the direct mapping of auditory signals into label sequences without the need for properly thought-out intermediary stages. Furthermore, the output does not require posterior processing. In contrast to the HMM-based model, the previous differences give end-to-end large vocabulary continuous speech recognition (LVCSR) the following characteristics: One network is created by combining many modules for collaborative training. One advantage of combining numerous modules is that the mapping between different intermediate states may be realized without requiring the creation of additional modules. Through joint training, the end-to-end model may search for globally optimum outcomes by using a function that is highly relevant to the final assessment criteria as a global optimization target. Instead of requiring additional processing to achieve true transcription or enhance recognition performance, it maps the input acoustic signature sequence directly to the text result sequence. This is in contrast to HMM-based models, which typically have an internal representation for a character chain's pronunciation. However, character level models are also recognized in certain conventional systems [3].

3. Applications of ASR

3.1. Application in house

ASR has achieved good results in-home use. Many household TVs have artificial intelligence voices. Smart voice similar to Siri, Xiao Ai classmate, etc., can achieve a series of operations through the host's conversation, such as turning on the TV, fast-moving retreat, next episode, and so on. Many operations need to be implemented through the remote control, and now only one sentence can be achieved. At the same time, smart homes are also a major application scenario. Through comprehensive wiring, network communication, security prevention, automatic control, ASR and other technologies, the house-related facilities related to the house are integrated, and the management system of high-efficiency residential facilities and family scheduling affairs will be established. Smart homes can improve home furnishings, environmental protection and energy saving, comfort and convenience. Common smart home devices include smart audio, smart lamps, refrigerators, air conditioners, etc. The emergence of smart homes makes people's lives more convenient and intelligent and also brings new development opportunities to the home industry.

3.2. Application in robots

ASR technology can be widely used in smart robots. The most prominent application is to use the navigation work of the robot. It can allow the robot to accurately avoid obstacles in various difficult environmental situations through voice dialogue and continue to perform the required work. For example, in forestry detection, the size of a tree is measured, and the robot needs to conduct a multi-directional observation and inspection on the potholes ground. Through ASR technology, you can use simple instructions such as walking left and right to control it to control. The robot avoids obstacles and stops at the specified observation point to observe the trees. At the same time, ASR technology can design the human-computer interaction interface. For example, the intelligent tour guide system of the robot can explore the attractions through dialogue; and the robot's smart takeaway system has a wide range of applications in today's hotels. People can use voice dialogue to send the robot to the designated room, takeaway households, and takeaways. What needs to be done is to place the takeaway on the robot and perform a conversation, which saves a lot of time and energy.

DOI: 10.54254/2755-2721/65/20240493

4. Improvements measures to improve ASR

4.1. Measures to improve immunity to interference

Most of the existing voice recognition systems can only work in a quiet environment. When the speaker is working in a noisy environment, emotional or psychological changes lead to pronunciation distortion and speed of pronunciation and the tone changes produce the LOMBARD effect or the LOUD effect. Because of the characteristics of the Chinese language, the processing of language information in Chinese is more difficult and complex than that of Western languages. The main manifestations are: the large character set of Chinese affects the rapid input of Chinese characters; The indistinguishability of Chinese words makes word segmentation a unique and primary problem in Chinese language understanding and processing. A large number of homophones and words make speech recognition difficult; Flexible and free language expression is difficult to express by the expression of Chinese language knowledge [4].

The purpose of the signal space algorithm is to eliminate the noise components in the noise-containing voice signal, so as to obtain an estimate of the pure voice signal. At present, there are mainly some methods. The fatigue of noise voice in hearing and improves the quality of voice communication. Now it is often used as the front-end processing process of anti-noise voice recognition. Because it improves the signal signal-to-noise ratio, it reduces the signal space caused by the noise. Spectral subtraction (SS) is a very small enhancement algorithm because the power of the background noise and voice can be similar to the addition of the addition. In the total power spectrum of the voice, you can obtain an estimate of the pure voice power spectrum of the frame, and the phase of the pure voice is replaced by the phase containing the noise voice with this frame. Finally, the noise can be greatly suppressed to enhance the voice [5].

At the same time, you can also use a password mode to ensure anti-interference, that is, we give a name for the module, only after calling his name can you talk to him; you need to pay attention to the following 2 points when the name is named: 3-6 words; pronunciation usually does not hear it often. What's more, we can use the "garbage words", like our keywords but the keywords that do not operate it. Select words similar to keywords, we do not process these words after recognition, that is, play mute or do not do any operation. Then these words are what we call the "garbage words", because it does not play any role after recognition, only to avoid interference keywords.

4.2. Errors detection and correction

When using the ASR system, translation errors and receiving errors will inevitably occur, and this requires system detection and correction. How to detect and correct errors on the premise of unmanned intervention has become a big problem today.

The purpose of error detection is to use characteristics produced by the ASR system, such as confusion network density, language model, and confidence scores, to ascertain whether a transcription error has occurred. Later on, a hypothesis word will be classified using those traits into one of two classes: either a valid term or an error [6].

However, there are no techniques that can be corrected 100 % for the time being. At present, there are many ways to correct the results of voice recognition, but there is no mainstream method. Most ASR post-amended studies are statistical methods based on the probability information of word recognition results. On the whole, the error correction can be divided into three major steps. The first step, the main purpose of this stage is to determine whether the text has an error to be corrected, and if it exists, it is passed to the latter two layers. The second step is to generate a correction candidate. The main purpose of this stage is to use multiple methods to make error correction candidates for the need to correct the wrong sentences. With this step, the overall error correction is more efficient. The third step is to evaluate the candidate for correction. The main purpose of this stage is that on the basis of the previous stage, the function uses the function to use screened error correction results to select the error correction result that is most in line with global needs. The highest person is the final error correction.

DOI: 10.54254/2755-2721/65/20240493

5. Conclusion

This paper mainly introduces the historical development of ASR technology, the application of ASR technology, and the development trend of ASR technology in the future. Generally speaking, speech recognition technology is one of the current hot technologies, after a hundred years of continuous development, the continuous innovation of basic theories, coupled with the abundance and progress of today's modern technology, has tended to be perfect, and has a good application prospect in various fields. This paper does not discuss the content of the basic ASR technology in depth, and the various basic modeling techniques have not been analyzed in detail, and this aspect still needs to be improved. As an emerging technology, it has entered millions of households, and it will also occupy a large part of the future. Not only for the convenience of home life but also to a higher technical level. In the era of intelligence, efficiency, and automation, it will also follow the overall trend of the times and progress to a more complete aspect.

References

[1] Juang, B. H., & Rabiner, L. R. (2005). Automatic speech recognition–a brief history of the technology development. Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara, 1, 67.

[2] Miao Miao & HaiWu Ma.(2006). Application of HMM in Automatic Speech Recognition System. Modern Electronics Technique(16),64-66.

[3] Wang, D., Wang, X., & Lv, S. (2019). An overview of end-to-end automatic speech recognition. Symmetry, 11(8), 1018.

[4] XiangZhi He.(2002).The Research and Development of Speech Recognition. Computer and Modernization(03), 3-6.

[5] Pei Ding.(2004). Noise Robust Technologies in Speech Recognition(Dissertation Submitted to Tsinghua University in partial fulfillment of the requirement for the degree of Doctor of Engineering).https://kns.cnki.net/kcms2/article/abstract?v=2F6201taHdcUwBjYSMP8SjmKHOTnRx5SwH8_3kv5Ng_nb-S1Vu5Y8YfFRVyK7Po26Yco0xAnHYKrZsZxBOMOSG4LHFw0xe5qR9xk5JnrqZUFNtOPQWGjNSfWqVPafDKR&uniplatform=NZKPT&language=CHS

[6] Errattahi, R., El Hannani, A., & Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: A review. Procedia Computer Science, 128, 32-37.