Research on speech recognition and its application in language disorders

Kangbo Wei

School of Biomedical Engineering, Tianjin Medical University, Tianjin, China
weikangbo6604@tju.edu.cn

Abstract. Since ancient times, language has been a fundamental medium for human communication and the expression of thoughts. The advancement of speech recognition technology has significantly enhanced the efficiency of generating, transmitting, storing, and acquiring speech information, thereby facilitating more convenient human-computer interaction. This technology is crucial in advancing societal progress by striving to develop machines that can understand and respond to human language naturally and empathetically. Moreover, speech recognition technology has revolutionized various fields, including healthcare, education, and customer service, by enabling real-time transcription, language translation, and voice-activated assistants. Despite the substantial progress made, the technology continues to evolve, particularly with the integration of deep learning and emotion recognition, paving the way for more sophisticated and empathetic human-computer interactions. These advancements aim to create a seamless and intuitive communication experience, bridging the gap between human and machine interactions and enhancing the overall user experience. In conclusion, the ongoing development of speech recognition technology promises to unlock new potentials and applications, further embedding itself as an indispensable tool in our daily lives.

Keywords: Speech recognition, emotion recognition, algorithm research, deep neural network, HMM.

1 Introduction

Since ancient times, people have used language to express their thoughts. Language is also the most effective and convenient way for people to communicate with each other. Through progressive research on speech recognition technology, scientific researchers have made the generation, transmission, storage and acquisition of speech information more efficient. Therefore, establishing a more convenient human-computer interaction method between humans and computers is of great significance to promoting the development of society. As the main technology for human-computer interactive communication, speech recognition has greatly promoted the scientific and technological progress of society. Its ultimate goal is to develop a machine that can understand human language and give appropriate feedback. This machine should not be limited to converting all words into written text accurately, but should be empathetic

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_55

and understand it more fluently. the meaning it represents. However, today's speech recognition technology is gradually unable to meet the demand. Currently, this technology has a wide range of applications in computers, communications and other fields.

This article mainly introduces the development status of speech recognition technology from three aspects:

Advancements in speech recognition technologies.

Speech recognition algorithms and applications;

Research on existing speech recognition algorithms and their advantages and disadvantages.

Emotion classification in speech recognition.

2 The development history of speech recognition

2.1 Origin and early development of technology

The 1950s saw the start of the development of speech recognition technologies. Based on the template matching algorithm, speech recognition technology was born. As early as 1952, the world's first isolated word recognition system was developed by the Bell Labs team led by Davis. Afterwards, IBM Labs also developed an isolated word recognition system in 1962. During this time, voice recognition was mostly accomplished by comparing speech signal parameters with pre-existing templates, which favored speech signal classification over in-depth recognition. During this time, individuals started to concentrate on the fundamentals of speech recognition technology, and feature extraction and cepstrum analysis technology developed rapidly during this period. After the 1960s, Soviet scientists proposed end-to-end detection technology, which can eliminate non-speech segments and retain the actual speech segments. This is an epoch-making breakthrough in speech recognition.

2.2 Developments from the 1970s to the present

The real start of speech recognition can be traced back to a series of research based on the GMM-HMM algorithm in the 1970s. Several important results during this period had an important impact on the field of speech recognition. For example: Linear Predictive Coding (LPC) technology solves the problem of feature extraction of speech signals; Dynamic Time Warping (DTW) technology time warps speech signals, solving the problem of speaking [1, 2]. The problem of unequal length of speech features caused by factors such as different speaking speeds of people. The Hidden Markov Model (HMM) is a statistical model used to describe the Markov process with hidden unknown parameters [3]. With its application, the accuracy and stability of speech recognition have been greatly improved. Since then, speech recognition has gradually become a research based on statistical probability model algorithms. Entering the 1980s, because the HMM model can not only describe the statistical characteristics of the speech signal, but also describe the transition from each short-term stable segment of the speech signal

to the next short-term stable segment. Therefore, this model is widely used in speech recognition systems. After that, Kaifu Li and others developed the SPHINX system based on the Gauss Mixed Model (GMM) framework [4]. It is the first-ever high-performance person-neutral, large-vocabulary speech recognition system. This model is currently mainly used to build phoneme-level acoustic models.

After the 21st century, speech recognition technology is gradually based on deep learning. The recognition accuracy of the HMM algorithm, which was once the core algorithm of speech recognition technology, gradually cannot meet the needs of practical applications, and the recognition rate increases slowly. So in 2009, Hinton applied the DNN (Deep Neural Network, DNN) network to the acoustic model of speech recognition for the first time to replace the output state modeling of GMM-HMM, that is, he proposed a new framework DNN-HMM [5]. Start a revolution in speech recognition using deep learning. Since then, strong recognition performance and system stability have been attained by hybrid recognition systems and end-to-end recognition systems made of common networks like recurrent and convolutional neural networks. Thus far, researchers both domestically and internationally have concentrated on neural network-based voice recognition systems.

3 Research on existing speech recognition algorithms and their advantages and disadvantages

3.1 Brief composition of speech recognition system

Speech recognition systems essentially include three parts: feature extraction, template recognition and matching, and reference template library.

Usually, the speech recognition process is roughly divided into two steps: first, a basic acoustic model and a language model for language analysis are established based on the extracted acoustic feature learning results; second, the characteristics of the input speech signal are compared with the characteristic parameters in the template library. Compare to find the template parameters that are most similar to the selected modeling method and obtain the recognition result.

In the specific process of speech recognition, signal preprocessing must first be carried out. Briefly, preprocessing includes operations such as prefiltering, sampling, analog-to-digital conversion, preemphasis, and frame windowing. Among them, the goal of pre-emphasis is to improve the spectrum of the high-frequency part, smooth the connection between the high-frequency part and the low-frequency part, and then solve the spectrum. Frame-based windowing allows the signal to have short-term stationary characteristics within a certain short period of time. A short-term stable signal is obtained, and then the traditional signal analysis method is used for subsequent processing.

The acoustic model and language model of the speech recognition system are particularly important in the speech recognition process. The acoustic model is the probability of generating a speech waveform after a given model. Its input is the feature vector sequence obtained after feature extraction of the speech signal. Currently, speech

recognition. The research on acoustic models in China is mainly about neural networks, and the synthesis and improvement of classic network structures for different application scenarios and requirements. The purpose of the language model is to predict the probability of character sequence generation and determine whether a language sequence is a normal sentence.

3.2 Research on speech recognition algorithms such as DTW and GMM

The important issues in speech recognition are the training of the speech template library and the template matching and recognition process. Speech recognition is fundamentally a process of pattern recognition. Moreover, in speech recognition, due to differences in speech between people, even the same sentence spoken by the same person at different times will have different characteristics. Therefore, directly comparing the reference template with the input template is very problematic. In order to solve this problem, the speech signal needs to be time warped (dynamic time warping, DTW). As a nonlinear method, this technology combines distance calculation with time warping. It is a relatively successful matching algorithm in isolated word speech recognition [6]. Dynamic time warping will decompose a problem into several small problems to solve. Then make judgments on each small issue.

In principle, the DTW algorithm non-uniformly distorts or bends the time axis of the speech signal to be measured, and then aligns the speech features to be measured with the template features, and then searches for a match with the minimum distance of the corresponding vector between the two templates. Path, according to the above steps, the regular function is obtained, which satisfies the condition of minimum cumulative distance when the vectors in the two templates are matched. It can be seen that DTW combines time warping and distance measurement to process speech signals. Generally speaking, the time warping (DTW) classifier will have good recognition results in isolated words and small and medium vocabulary recognition. Compared with other algorithms, DTW has fast recognition speed and low system overhead. It is excellent in speech recognition. algorithm. However, when faced with a large number of words and non-specific speech recognition, the recognition results will have many problems. At this time, the recognition effect of speech signals using Hidden Markov Model (HMM) will be significantly improved.

Baum and others established the theory of Hidden Markov Models in the 1970s, and then Rabiner and others from Bell Labs studied it in detail. Today, HMM is still an important algorithm in the field of speech recognition. Hidden Markov Model (HMM) consists of two random processes: the first is the Markov chain, which is used to describe the transition of states and is a basic random process; because the HMM model is observed from the perspective of an observer, Therefore, only the observed values can be observed, but the runtime state and its transformation relationship cannot be directly observed. Therefore, there must be a random process to "perceive" the characteristics of the state. Then another random process appears. It is used to describe the corresponding statistical relationship between observations and states. This algorithm is therefore called a "hidden" Markov model.

The two most often used algorithms in voice recognition at the moment are Dynamic

Time Warping (DTW) and Hidden Markov Model (HMM). In comparison, the theory of the DTW algorithm is relatively easy to understand. However, for non-specific people, large vocabulary, and continuous word recognition, the recognition effect of HMM is significantly better than that of DTW, but its theory is more complex and requires a lot of training to obtain the reference template.

3.3 Optimization and improvement of the current algorithm

The general DTW algorithm will save all frame matching distance matrices and cumulative distance matrices and search for all pathways that satisfy the slope constraint requirements inside the boundary range in order to determine the optimal path throughout the template matching phase. This procedure is going to cause more issues. large-scale activities. This in turn uses up a lot of system resources and somewhat slows down the system's recognition speed. Regarding the aforementioned issues, researchers have suggested modifying the conventional DTW algorithm in the following ways to reduce the quantity of system calculations and increase operating speed: First, limit the path function to a parallelogram. Each time you search forward one frame on the x-axis, you only need to calculate the cumulative distance of the previous column. Therefore, you only need to apply for two column vectors in the program: one is used to save the current column cumulative distance. The column vector d of the distance, and the other is the column vector D used to save the cumulative distance of the previous column (D in the cumulative minimum distance formula represents the cumulative distance, d represents the frame matching distance), and there is no longer a need for a matrix to save the entire distance, so Reduce the amount of system storage and calculations [7].

4 Speech recognition and emotion recognition based on deep learning technology

4.1 Mainstream research directions in emotion recognition

To enable computers to accurately discern the connection between emotion measures and speech information carriers is the goal of speech emotion recognition. Speech emotion recognition is a key area of research in artificial intelligence and plays a significant role in human-computer interaction. Over three decades have been spent in research on voice emotion recognition. As early as 1995, Professor Picard of MIT first proposed "Affective computing" and established "Affective computing" as a new discipline in the field of computing [8]. Overall, speech recognition techniques can be broadly classified into two groups: end-to-end and speech emotion recognition approaches based on speech emotional qualities. The generic speech emotion detection approach primarily uses the speech signal's acoustic properties to identify emotions in speech signals. For this reason, one crucial area of research in speech emotion identification is the extraction of emotion-related acoustic characteristics. Researchers are currently replacing conventional acoustic features with deep learning technology to

achieve end-to-end speech emotion recognition by performing non-linear processing of the original voice signal to build a deep representation of the data. So far, speech emotion recognition has become an important part of the national artificial intelligence development strategic layout.

4.2 Models for emotion representation and how they work

Objective analysis shows that human emotions are characterized by ambiguity, subjectivity, complexity, and time-varying characteristics. Discrete emotion theory and dimensional emotion theory are the two main categories into which emotion representation techniques now fall. Discrete emotion theory uses discrete states to describe emotions, such as excitement, anger, sadness, disgust, etc. Discrete emotion theory does not use highly professional judgment tools very frequently. Further down, it can be divided into basic emotion categories and complex emotion categories. Emotion is defined at a place in a different dimensional space by the dimensional emotion theory, which describes emotion using continuous dimensional space. The angle and separation between points in the space affect how similar and distinct various emotions are from one another. Articulate [9]. With the assistance of the SAM system, it quantifies the PAD model values, and FEELTRACE quantifies the VA model dimension values. ANNEMO labels one dimension at a time, and its results are highly accurate [10].

4.3 Emotion database for speech recognition

The quality of the database directly impacts speech emotion recognition performance. The speech emotion database is therefore essential for speech emotion recognition. This phenomenon is very prominent in popular data-driven model algorithms. The existing emotional voice libraries mainly include performance type, guided type, and spontaneous type. Most of the speech emotion corpus in the speech recognition database comes from sharing by multimedia users. This corpus has greater feasibility and appeal. With the development of multimedia, speech emotion corpus will extract more content from platforms such as broadcasts and short videos. Examples you can see include the emotion corpus ShEMO, OMG, MSP-PODCAST, etc [11-13].

4.4 Existing speech recognition emotion recognition models

Two broad categories can be used to categorize emotion recognition models: Speech emotion identification methods based on speech acoustic features fall into the first group. These approaches typically take certain artificially created features from the speech signal and feed them to the classifier in order to finish the recognition. Commonly used speech emotion features are mainly divided into three categories: prosodic features, spectral features and tone quality features. On the basis of these features, Chen Yiling et al. used a combination of MFCC and spectrogram, and used SVR as a regression prediction model. Compared with MFCC alone, the performance of dimensional speech emotion recognition was significantly improved. In addition,

Akagi et al., a research team at the Hokuriku Advanced Institute of Science and Technology in Japan, believe that the emotional information in human perception of speech does not come directly from changes in acoustic features, but is conveyed by a combination of emotional semantics or adjective expressions [14]. As a result, the research team constructed a three-layer model of speech emotion based on auditory perception. The bottom layer is acoustic features, the middle layer is a combination of some adjectives, and the upper layer is emotional categories or dimensional emotional space [15]. Elbarougy et al. used this perceptual three-layer model and related the three layers through a fuzzy inference system to construct a simple speech emotion recognition model [16, 17]. This model has a higher recognition rate than the two-layer model that directly maps acoustic features to emotions. high.

Based on end-to-end speech emotion recognition is the other kind. Speech emotion recognition uses a wide range of end-to-end learning framework models these days. By eliminating the need for manual feature extraction and better adjusting to emotion recognition tasks, it uses a deep neural network to automatically learn an emotional representation form from the original speech data that contains emotional information, two-dimensional spatial information, and temporal context information. Raw speech data contains rich information. Using mathematical models to represent it as speech acoustic features may cause the omission of emotional information. Using an end-to-end approach can capture more complete emotional information directly from the original signal.

5 Application of speech recognition to language barrier problems

5.1 Stuttering correction speech synthesis system

Stuttering is a common language disorder, and as the application of artificial intelligence gradually enters the medical field, this problem has also been solved. Based on this, scientific researchers developed the Stuttering Voice Synthesis System (SVSS). Based on the characteristics of stuttering speech, this system extracts speech characteristic parameters, builds a parameter model library required for system development, and successfully develops an SVSS system that meets user needs. The system extracts the MFCC parameters in the signal source as the characteristic parameters of speech, uses the LPC analysis method for speech compression, establishes an HMM training model, uses the Viterbi algorithm to correct the spectral envelope parameters, and uses the k-means algorithm for optimization, so that the training model can output the best The best matching results are obtained, and finally the matching results are analyzed combined with the existing rhythm parameters. This study effectively applies intelligent speech technology to stuttering correction, which can greatly reduce the large amount of manpower and financial resources consumed in the stuttering correction process, and plays a role in promoting research on auxiliary treatment of stuttering [18].

5.2 The application effect of artificial intelligence voice signal collection and recognition method in patients with Alzheimer's disease

Alzheimer's disease (AD) is a progressive and irreversible neurodegenerative disease. The main symptoms of patients are memory and cognitive function decline, accompanied by language impairment and neurological dysfunction, researchers are currently trying to use speech recognition technology to diagnose Alzheimer's disease [19, 20]. The specific method is as follows: The first step is to establish a deep learning recognition model, use confirmed samples to conduct deep learning neural network model learning and training, and obtain a deep learning model. The second step is to collect voice signals from patients diagnosed with AD, and ask the subjects to read three set paragraphs of text, each of 60 to 80 words. The voice signals of all subjects are recorded with the same recorder. The third step is speech signal screening and recognition. The subject's speech signal is input into the deep learning recognition model, the speech signal is screened and recognized, and a conclusion is drawn.

Through this experimental method, it can be considered that the intelligent speech signal collection and recognition method is an effective and practical AD screening technology and has great potential in early diagnosis and intervention treatment of AD. This method can be combined with the traditional AD scale and brain

Spinal fluid testing and imaging tests complement each other.

6 Challenges Facing Speech Recognition

6.1 Speech recognition technology still faces the following challenges

Lack of an emotional speech database that has both scale and quality and can be used universally: Today's databases are large in number, small in scale, and not universal. Researchers need to choose different libraries according to their research goals, making it inconvenient to compare each other's recognition results.

Model compression and acceleration: At present, models with good speech recognition effects are often complex and large in scale, require redundant data resources, and are not suitable for mobile devices.

Speech recognition emotion corpus: Current deep learning methods require a large amount of data for training to accurately realize their functions. The current speech emotion data set has a small total corpus, which will reduce the robustness of the deep learning model.

7 Conclusion

Deep neural networks are currently the foundation of the majority of popular voice recognition techniques. These techniques fall into two main categories: the first involves using a specific neural network to replace a single module—such as feature extraction, acoustic modeling, or language modeling—in the conventional speech recognition method; the other involves implementing neural network-based end-to-end

speech recognition. Current speech recognition research has advanced to the point where business needs are met under ideal conditions, including quiet surroundings. However, there are a number of challenging cases in real-world applications, including professional language settings, minority language detection, far-field sound sources, and speaker accents. This makes the implementation of speech recognition applications in complex scenarios still facing challenges. In short, a wealth of research results have been achieved in the field of speech recognition, but there is still a long way to go.

References

1. Müller, M. Dynamic Time Warping. Springer, Berlin Heidelberg (2007).
2. Zhang, J., Qin, B. DTW Speech Recognition Algorithm Based on Optimized Template Matching. In: World Automation Congress 2012, pp. 1-4 (2012).
3. Rabiner, L. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 77(1), 257-286 (1989).
4. Lee, K. F., Hon, H. W., Reddy, R. An Overview of the SPHINX Speech Recognition System. IEEE Transactions on Acoustics, Speech, and Signal Processing, 38(1), 35-45 (1990).
5. Mohamed, A., Dahl, G., Hinton, G., Others. Deep Belief Networks for Phone Recognition. In: NIPS Workshop on Deep Learning for Speech Recognition and Related Applications, pp. 39 (2009).
6. Zhao, L. Speech Signal Processing. 2nd ed. Mechanical Industry Press, Beijing (2003).
7. Cheng, Y. Study on Speech Recognition Algorithms and DSP Implementation. Master's thesis, Anhui University of Science and Technology, Anhui, China (2015).
8. Picard, R. W. Affective Computing. The MIT Press, Massachusetts (2000).
9. Tao, J., Chen, J., Li, Y. A Survey of Speech Emotion Recognition. Signal Processing, 39(4), 571-587. DOI:10. 16798/j.issn.1003-0530.2023.04.001 (2023).
10. A Review of Dimensional Speech Emotion Recognition. DOI:10. 16661/j.cnki.1672-3791.2022.08.5042-0208 (2022).
11. Mohamad Nezami, O., Jamshid Lou, P., Karami, M. ShEMO: A Large-Scale Validated Database for Persian Speech Emotion Detection. Language Resources and Evaluation, 53(1), 1-16 (2019).
12. Barros, P., Churamani, N., Lakomkin, E., et al. The OMG-Emotion Behavior Dataset. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1-7 (2018).
13. Lotfian, R., Bousso, C. Building a Naturalistically Balanced Emotional Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings. IEEE Transactions on Affective Computing, 10(4), 471-483 (2017).
14. Chen, Y., Cheng, Y., Chen, X., et al. Speech Emotion Recognition in PAD Three-Dimensional Affect Space. Journal of Harbin Institute of Technology, 50(11), 160-166 (2018).
15. Huang, C., Akagi, M. A Three-Layered Model for Expressive Speech Perception. Speech Communication, 50(10), 810-828 (2008).
16. Elbarougy, R., Akagi, M. Speech Emotion Recognition System Based on a Dimensional Approach Using a Three-Layered Model. In: Proceedings of the 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 1-9. IEEE, Hollywood, CA, USA (2012).
17. Elbarougy, R., Akagi, M. Cross-Lingual Speech Emotion Recognition System Based on a Three-Layer Model for Human Perception. In: 2013 Asia-Pacific Signal and

Information Processing Association Annual Summit and Conference, pp. 1-10. IEEE, Kaohsiung, Taiwan, China (2013).
18. Kou, Y. Research and Application of Speech Synthesis Technology in Stuttering Correction. Master's thesis, Tianjin University, Tianjin, China (2012).
19. Wu, Y., Jiang, Y., Tang, L., et al. Screening Test for Early Alzheimer's Disease Using a Rapid Cognitive Assessment Scale. Chinese Journal of Mental Health, 34(2), 106-111 (2020).
20. Xing, T. Comparison of Two Commonly Used Scales, MMSE and HDS, for Alzheimer's Disease. Massage and Rehabilitation Medicine, 11(4), 16-17, 20 (2020).