Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ26545

Overview of Speech Recognition Algorithms and Their Applications

Xuyang Chen

School of Mechanical Engineering, Nantong University, Nantong, China
3368136462@qq.com

Abstract. Speech recognition technology, a pivotal element in human-computer interaction, has witnessed substantial advancements in recent years, propelled by the synergies of deep learning and big data. This paper provides a systematic review of the evolution of speech recognition algorithms, delineating the principal characteristics and application contexts of traditional speech recognition algorithms, such as Hidden Markov Models (HMM), deep learning-based algorithms, including Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN), and end-to-end speech recognition algorithms. Furthermore, this study delves into the multifaceted applications of these algorithms in domains such as voice assistants (e.g., Siri and Alexa), machine translation, and meeting transcription, elucidating their transformative impact. The paper also synthesizes the prevailing speech recognition technologies and the challenges they confront, with a particular emphasis on the limitations of commonly used language recognition algorithms, such as susceptibility to noise, accent variability, and data dependency. Through this comprehensive analysis, the paper aims to illuminate the current state and future trajectories of speech recognition technology. This paper identifies and summarizes the shortcomings of commonly used language recognition algorithms.

Keywords: Speech Recognition, Deep Learning, End-to-End Model, Voice Assistant, Machine Translation

1. Introduction

Automatic Speech Recognition (ASR) is the technology that converts human speech into text and is an important research direction in the fields of artificial intelligence and human-computer interaction. With the proliferation of mobile internet and smart devices, speech recognition technology has become increasingly widespread in daily life, such as in smart speakers, voice assistants, and autonomous driving. Traditional speech recognition systems relied on Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM). However, with the development of deep learning technology, neural network-based speech recognition algorithms have gradually become mainstream, greatly improving recognition accuracy and robustness [1]. This paper uses literature analysis to provide an overview of the main types of speech recognition algorithms, key technologies, and their applications, aiming to provide a reference for relevant researchers and practitioners.

© 2025 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: 10.54254/2755-2721/2025.BJ26545

2. Main types of speech recognition algorithms

2.1. Traditional speech recognition algorithms

Traditional speech recognition systems typically employ statistical model-based methods, primarily including Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM). HMM is used to model the time series characteristics of speech signals, effectively capturing the dynamic changes in speech; GMM is used to model acoustic features, describing the probability distribution of speech signals [2]. This method was widely used from the 1990s to the early 21st century, for example, in early telephone speech recognition and voice input systems. However, due to its reliance on manually designed features and assumptions about the statistical properties of speech signals, traditional methods face issues of low accuracy and poor robustness when dealing with complex speech environments (such as noise interference) and large-scale data.

2.2. Deep learning-based speech recognition algorithms

With the rapid development of deep learning technology and the continuous growth of computational resources, neural network-based speech recognition algorithms have emerged on a large scale since the early 2010s, completely reshaping the technological landscape of the field and significantly and continuously improving the overall performance of systems [3]. Deep Neural Networks (DNN), as powerful nonlinear function approximators, are revolutionary in their ability to automatically and end-to-end learn from raw speech signals (such as Mel-frequency cepstral coefficients (MFCCs) or filter bank FBank features) to more abstract and discriminative high-level feature representations [4]. This strong representation learning capability completely replaces the manual feature engineering part in the traditional Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) system, especially its core GMM acoustic model. By training on large-scale annotated speech data, DNN can capture finer nonlinear structures and contextual information in complex speech patterns, greatly improving the accuracy and robustness of acoustic modeling, and its performance improvement was widely regarded as a step-change breakthrough at the time.

However, speech is inherently a continuous and dynamic time series signal, and traditional feedforward DNNs have limited ability to model long-term dependencies when processing such sequences. To address this challenge, Recurrent Neural Networks (RNNs), which are more suitable for sequence modeling, were introduced, especially the improved Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). The RNN structure has an inherent "memory" characteristic that can handle input sequences of arbitrary length; LSTM, through its cleverly designed gating mechanisms (input gate, forget gate, output gate), effectively alleviates the common gradient vanishing/exploding problems in standard RNN training, enabling it to stably learn and capture very long-span speech context information. This capability is crucial for understanding continuous speech streams, distinguishing confusable phonemes (such as "b" and "p"), and handling coarticulation and tonal changes, making the LSTM-RNN architecture the preferred model for continuous speech recognition (ASR) tasks, especially large vocabulary continuous speech recognition (LVCSR) tasks.

Benefiting from increased model capacity and advanced training techniques, they exhibit stronger suppression capabilities in noisy environments (such as background music, traffic noise, and multi-speaker interference); in multilingual/dialect recognition, by learning feature representations shared across languages, systems supporting multiple languages/dialects can be built and scaled more efficiently [5]. Moreover, when faced with large-scale datasets, deep learning models not only utilize

DOI: 10.54254/2755-2721/2025.BJ26545

data efficiently but also continue to improve in performance as the amount of data increases. This unprecedented performance and adaptability have rapidly surpassed traditional methods, indisputably becoming the core engine of current mainstream speech recognition technology [6]. Industry-leading systems such as Google Voice Search and Baidu Speech Recognition Platform have deeply integrated and optimized multi-layer DNN, LSTM/GRU, and even Transformer deep learning models in their core acoustic and language modeling modules, driving the implementation of massive applications such as mobile voice input, smart home control, real-time caption generation, meeting transcription, and intelligent customer service [7].

2.3. End-to-End speech recognition algorithms

End-to-End (E2E) speech recognition is a new paradigm that has emerged and rapidly developed in recent years. Its core idea is to abandon the complex process in traditional speech recognition systems (Hybrid Systems) where acoustic models (AM), pronunciation lexicons (Lexicon), and language models (LM) need to be designed, trained, and integrated independently in stages. It aims to build a single, unified deep neural network model that can directly learn the mapping from input speech signals (usually spectral features such as FBank) to the final text output [8]. This paradigm shift greatly simplifies the overall system architecture, reduces reliance on domain knowledge and manually designed components (such as state tying, decision trees, pronunciation rules), significantly lowering system complexity and facilitating deployment and maintenance. Currently, the mainstream end-to-end model architectures are mainly divided into two categories: Connectionist Temporal Classification (CTC) and models based on the Attention Mechanism (and their variants and fusion models). The key innovation of the CTC model is the introduction of a special "blank" label, which allows the model to be trained without precise frame-level alignment when processing variable-length input speech frame sequences and output character sequences [9]. It achieves this by oversampling the input sequence in the time dimension (output path length equals the number of input frames) and allowing the output of blank and characters, then merging repeated characters and removing blanks during inference to form the final text. This method is particularly suitable for continuous speech recognition. Attention mechanism models, on the other hand, adopt a more flexible, dynamic soft alignment approach. The core component of the model (attention module) can "actively focus" on the most relevant parts of the input sequence (usually several consecutive frames) when generating each output character. This content-aware mechanism enables the model to more accurately capture the complex correspondence between phonemes-characters or subword units-characters, often surpassing CTC models in recognition accuracy (especially in scenarios with large vocabularies and high grammatical complexity), particularly when combined with powerful encoder-decoder architectures like Transformer.

However, end-to-end models also face challenges, mainly manifested in their heavy reliance on large-scale annotated training data (typically requiring thousands of hours of speech-text paired data to achieve optimal performance) and the demand for powerful computational resources (especially GPU memory) [10]. The advantages of the end-to-end approach have shown great potential and commercial value in several cutting-edge applications. For example, in intelligent voice assistants (such as Siri, Alexa) for wake-word and command recognition, high-precision real-time speech transcription services, high-quality real-time speech translation systems, and video caption generation combined with multimodalities, end-to-end models are increasingly becoming the mainstream choice for industrial deployment due to their simplicity and continuously optimized performance. Their development has also accelerated from early RNN/LSTM+Attention to backbone architectures like Transformer/Conformer (Convolution-augmented Transformer), and

DOI: 10.54254/2755-2721/2025.BJ26545

continues to explore technical directions such as RNN-Transducer (RNN-T), Neural Transducer, and streaming attention variants to meet the stringent requirements of real-time streaming recognition [11].

3. Applications of speech recognition algorithms

3.1. Voice assistants

Voice assistants are undoubtedly one of the most widespread and influential applications of speech recognition technology today, almost becoming a standard feature of smart living. Mainstream products represented by Apple's Siri, Amazon's Alexa, Google Assistant, and domestic ones like Baidu Xiaodu and Alibaba Tmall Genie deeply integrate speech recognition technology as their core interaction entry point. These voice assistants use advanced speech recognition modules to capture and transcribe users' voice commands in real time, converting them into text information. The subsequent Natural Language Understanding (NLU) module is responsible for deeply parsing the user's intent (Intent) in the text instructions and extracting key semantic slots (Slots), thereby triggering the dialogue management system to perform corresponding operations or call services [12]. This covers a wide range of scenarios from daily functions (such as playing music, checking the weather, setting reminders, controlling smart home devices) to complex interactions (such as multi-turn dialogue for flight information, online shopping, or restaurant booking).

The comprehensive penetration of deep learning technology, especially the application of end-to-end architectures, is key to driving the leap in capabilities of voice assistants. At the front end of speech recognition, deep neural networks (such as end-to-end models based on Transformer or Conformer) significantly improve the real-time performance and accuracy of recognition, especially when deployed lightweight on devices, enabling quick response to wake words (such as "Hey Siri") and execution of immediate commands. At the core intent understanding layer, deep learning-based NLU models (such as pre-trained language models like BERT, XLNet) can more finely handle colloquial expressions, understand context, disambiguate, and even capture users' implicit needs by learning from massive dialogue corpora. At the dialogue management level, models based on reinforcement learning or sequence-to-sequence (Seq2Seq) can optimize dialogue strategies to achieve more natural and smooth multi-turn conversations.

The popularity and performance improvement of voice assistants complement each other, greatly enhancing the naturalness, convenience, and efficiency of human-computer interaction. Users can complete operations just by speaking, without the need for keys or touch screens, greatly expanding the possibilities of interaction, especially in scenarios where hands and eyes are restricted (such as driving, kitchen work) or multitasking. This directly promotes the vigorous development and widespread implementation of the smart home ecosystem (IoT), allowing users to centrally control numerous smart devices such as lights, air conditioners, security, and entertainment systems through voice assistants. Its application scenarios have extended from mobile phones to diverse hardware platforms such as smart speakers, wearable devices, smart TVs, and in-vehicle infotainment systems (IVI), becoming a key interaction node in the era of the Internet of Things [13].

3.2. Machine translation

Speech recognition technology plays a core and indispensable role in modern machine translation systems, especially in real-time speech translation scenarios that pursue seamless communication. The typical process of such systems highly relies on speech recognition (ASR) as the crucial first

DOI: 10.54254/2755-2721/2025.BJ26545

step: first, the system captures the source language (such as English) speech signal through a high-quality microphone array and uses robust noise reduction and acoustic processing techniques for front-end processing of the signal. Then, an advanced speech recognition engine (usually based on end-to-end models such as Transformer or Conformer) decodes the processed speech in real time, accurately transcribing it into an intermediate text representation of the source language. The accuracy of this step is crucial because it is the foundational input for the entire translation process.

After obtaining the source language text, the system calls upon powerful Neural Machine Translation (NMT) engines. These engines generally use deep sequence-to-sequence models based on the Transformer architecture, which, by learning from massive parallel corpora, possess a deep understanding of contextual semantics and can generate more natural expressions that conform to the habits of the target language (such as Chinese). Finally, natural speech synthesis technology (TTS) converts the translated target language text into smooth speech output in real time, completing the closed loop of cross-language communication. An intuitive application example is the voice function of Google Translate: users speak English into their phones, and the system outputs clear Chinese speech (and text) almost synchronously, greatly facilitating travelers and business people.

The breakthrough of end-to-end speech recognition models has revolutionary significance for improving the performance of real-time speech translation. Such models significantly reduce the processing latency from speech input to text output by simplifying the system architecture and directly optimizing the mapping relationship, while the model's strong representation learning capability also greatly reduces the error rate of speech transcription [14]. This makes speech translation services truly practical and efficient, with their value fully demonstrated in fields such as international conferences, multinational business negotiations, multilingual tourism guidance, emergency rescue collaboration, and remote education [15]. In these scenarios, the immediacy, accuracy, and privacy of communication (compared to public translation) are all core requirements.

3.3. Meeting transcription

In modern collaboration and dissemination scenarios where efficiency is paramount, speech recognition technology plays a key role as a productivity tool, especially in the two core areas of automated meeting transcription (Meeting Recaps) and real-time caption generation (Live Captioning). These applications can convert ongoing spoken content into structured text in real time and automatically, greatly freeing up manpower (no longer needing dedicated stenographers), reducing recording costs, and significantly improving information accessibility. For people with hearing impairments, non-native participants, or users working in noisy environments, real-time captions provide crucial accessibility support.

With technological advancements, deep learning-based speech recognition systems have become the cornerstone of these scenarios. Their core advantage lies in their ability to effectively handle complex real-world acoustic environments [2]:

A. Multi-speaker Recognition: Advanced end-to-end models (such as those based on Transformer architectures) can better model the speech characteristics and speaking styles of different speakers. B. Overlapping Speech Handling: By introducing mechanisms like Multi-Head Self-Attention and special training objectives (such as Permutation Invariant Training, PIT), the model can attempt to separate and recognize multiple sound sources speaking simultaneously, although this remains a highly challenging task.

C. Noise & Reverberation Robustness: Combining the powerful feature extraction capabilities of deep neural networks (such as the Conformer model that integrates CNN for capturing local features

DOI: 10.54254/2755-2721/2025.BJ26545

and Transformer for modeling global dependencies) with multi-channel signal processing (microphone arrays) and data augmentation strategies (adding various noise and reverberation simulations during training), the system can maintain high recognition rates under various acoustically degraded conditions.

4. Conclusion

This paper mainly explores different technologies of speech recognition, including traditional speech recognition algorithms, deep learning-based speech recognition algorithms, end-to-end speech recognition algorithms, and the application scenarios of speech recognition technology. Speech recognition technology has made significant progress over the past few decades, especially with the introduction of deep learning and end-to-end models, which have greatly improved recognition accuracy and robustness. From traditional HMM-GMM systems to modern neural network models, speech recognition algorithms have undergone a transformation from statistical modeling to data-driven approaches. However, current technology still faces challenges in handling dialects, low-resource languages, and extreme noise environments. This paper has not delved deeply into the principles of speech recognition technology. Future research directions include: a. Developing more efficient model architectures to reduce computational resource requirements; b. Exploring unsupervised and self-supervised learning methods to reduce reliance on annotated data; c. Improving the performance of multilingual and cross-lingual speech recognition; d. Enhancing the security and privacy protection of speech recognition systems.

References

[1] Jiang Yinhe. Research on Multi-Modal Emotion Recognition Algorithms Based on Video and Speech [D]. Hangzhou Dianzi University, Electronic Information, 2024: 1-9.
[2] Fan Peng. Research on Key Technologies for Speech Recognition in Mixed-Language Air Traffic Control [D]. Sichuan University, Computer Science and Technology, 2024: 5-7.
[3] Chen Shuo. Research on the Application of Deep Learning Neural Networks in Speech Recognition [D]. South China University of Technology, School of Electronics and Information, 2013: 28-31.
[4] Ji Xueying. Research on Speech Recognition Technology Based on Deep Learning [D]. North China University of Technology, Information and Communication Engineering, 2024: 15-19.
[5] Zhu Shuqin. Research on Key Technologies of Language Recognition Systems [D]. Xidian University, Computer Application Technology, 2004: 8-15.
[6] Zhao Xiaoqun, Zhang Yang. Review on the Construction of Acoustic Models for Speech Keyword Recognition Systems [J]. Journal of Yanshan University, 2017, 41(6): 471-478.
[7] Lu Lin, Wang Dong. A Brief Discussion on the Development Trends of Sound Recognition Models [J]. Automobile Applied Technology, (12): 186-188.
[8] Shang Kailun. Research on Federated Learning Algorithms for Data Heterogeneity and System Heterogeneity Problems [D]. Beijing University of Chemical Technology, Electronic Information, 2024: 14-18.
[9] Lin Quan. Research on End-to-End Chinese Language Recognition Algorithms [D]. Southeast University, Electronic and Communication Engineering, 2022: 23-39.
[10] Jin Xiuli. Research and Implementation of End-to-End Language Recognition Algorithms [D]. Lanzhou Jiaotong University, Information and Communication Engineering, 2023: 27-32.
[11] Qing Yuan. Research on the Engineering Application of Speech Recognition Based on RNN-T End-to-End Scheme [D]. Nanjing University of Posts and Telecommunications, Software Engineering, 2022: 22-28.
[12] Li Xue'ao. A Brief Introduction to Intelligent Speech Interaction Control Technology for Virtual Digital Humans [J]. China Equipment Engineering, 2023, 12(1): 28-30.
[13] Wang Bei, Zhao Ruisong, Niu Yiru, Guo Yuanyuan, Zhao Dongyang, Deng Yunfeng, Zhong Dingrong. Application and Exploration of Intelligent Speech Recognition Technology in Pathology Department [J].
[14] Li Zhen. Research on End-to-End Neural Network Machine Translation Technology [D]. Information Engineering University of Strategic Support Force, Information and Communication Engineering, 2020: 4-11.

DOI: 10.54254/2755-2721/2025.BJ26545

[15] Shi Xiaohu, Yuan Yuping, Lv Guilin, Chang Zhiyong, Zou Yuanjun. Review of Model Compression Algorithms for Automatic Language Recognition [J]. Journal of Jilin University, 2024, 62(1): 122-131.