Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/121/2025.19493

Evolution and Challenges in Speech Recognition Technology: From Early Systems to Deep Learning Innovations

Jingjia Xu[1,a,*]

[1]*Merrill College, Jack Baskin Engineering, University of California Santa Cruz, Santa Cruz, 95064, United States*
a. jingjiaxu0204@gmail.com
**corresponding author*

Abstract: Speech recognition technology is user-friendly and enables machines to understand and process human language, converting spoken language into text. As a critical component in numerous applications, this technology facilitates natural, hands-free interaction, enabling individuals to communicate and operate devices seamlessly, thereby enhancing the convenience and accessibility of everyday life. Additionally, speech synthesis assists users in multitasking and offers benefits to the visually impaired. Translation applications enable users of different languages to communicate with each other through one-to-one language conversion in the program. Speech recognition technology has evolved from rule-based methods to modern deep learning models. This paper explores the development history of speech recognition systems, focusing on analyzing its key technical milestones and challenges. Through a combination of historical analysis and technical insights, this paper examines how algorithms such as deep learning and neural networks can significantly improve speech recognition accuracy. The paper concluded that while deep learning has significantly boosted performance, hurdles such as managing diverse accents and environmental noise persist, indicating that there is still potential for future advancements.

Keywords: speech recognition, algorithms, deep learning, language, machine learning.

1. Introduction

Speech recognition technology leverages artificial intelligence and machine learning to integrate grammar, syntax, structure, and semantics, thereby enhancing language processing accuracy. Voice input is more convenient and faster and allows users to interact with virtual assistants through natural voice and complete tasks intuitively and unobstructedly. The diversity of human languages and pronunciations poses significant challenges to speech recognition. Speech recognition technology spans linguistics, mathematics, and statistics and is regarded as one of the most complex technologies in computer science. Differences in pronunciation, accent, and intonation between groups significantly affect recognition accuracy. This article will further explore the development history and critical challenges of speech recognition technology, analyze the evolution of algorithms such as deep learning and neural networks, and look forward to future improvements.

© 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: 10.54254/2755-2721/121/2025.19493

2. Origin and development of speech recognition technology

Natural language processing is mainly used in artificial intelligence for human-computer communication through voice and text. Many mobile devices, such as text messaging, have this system installed for more accessible communication and dialogue. The hidden Markov model is based on the Markov chain model, which stipulates that the probability of a given state depends on the current state, not its previous state[1]. This model allows hidden events to be incorporated into the probability model for sorting and labeling. N-gram is the simplest language model for assigning probabilities to sentences or phrases. Neural networks are mainly used in deep learning algorithms to process training data by simulating the inter connectivity of the human brain through node layers. Neural networks learn this mapping function through supervised learning and adjust it according to the loss function through gradient descent. The speaker classification algorithm recognizes and segments speech by speaker identity. It helps the program better distinguish between individuals in a conversation.

Since the 1950s, researchers have been studying and creating speech recognition technology. Almost every decade has a breakthrough invention in speech recognition technology. In 1952 the first automatic speech recognition machine, "Audrey," was launched[2]. This giant machine invented by Bell Labs can recognize the basic units of speech, namely phonemes. It can recognize any number spoken by humans from 0-9. Her creation hopes to automate the work of manual operators or replace some operators' initial or transfer work. However, Audrey was a large machine that consumed a considerable amount of electricity. Audrey is like a human fledgling; she immensely likes her creator, HK Davis. When recognizing HK Davis's spoken numbers, the accuracy rate can exceed 90%. However, the accuracy rate for other testers will drop to 70-80%. Perhaps the accuracy rate would drop for unfamiliar voices, but the computer system was inflexible then, and the calculation speed was slow. This historical context helps us appreciate the evolution of speech recognition technology, from the early days of Audrey to the sophisticated systems today. Audrey was an early invention that appeared before the invention of general-purpose computers.

At the 1962 World's Fair, IBM demonstrated the "Shoebox" machine, which could understand 16 spoken English words[2]. The shoebox is a machine used to recognize speech. Most of this system is based on template matching, matching individual words with stored speech patterns. Based on template matching, the Shoebox was a compact device, about the size of an actual shoebox, with an advanced circuit of 31 transistors, each capable of recognizing words and sounds[3]. When operating, the experimenter needs to speak into the shoe box's microphone, and the shoebox will then convert the sound into electrical impulses and classify them before calculating the answer. The shoebox is small, about the size of an actual shoebox. Its circuit is very advanced, consisting of 31 transistors. Each of these 31 transistors can recognize less than two words but recognize complete words plus sounds. In contrast, other early speech recognition machines required up to 200 transistors.

After analyzing the design choices of two earlier speech recognition systems, the Hearsay-1 and Dragon systems, Carnegie Mellon University developed the Harpy speech recognition system[2][4]. Hearsay-1 used programs to represent knowledge, while the Dragon system relied on Markov networks. After comparative studies, Harpy uses finite state transition networks to represent knowledge and determines the optimal path by searching several "best" paths in parallel, thereby reducing the number of state probability updates. Harpy introduces new techniques to improve performance and speed, including folding common sub-networks to reduce complexity, avoiding complete phoneme matching at each sampling, and semi-automatic learning of vocabulary representations and phoneme templates. In addition, it handles inter-word phenomena by applying connection rules when the network is generated, avoiding the time-consuming application of phoneme rules during the recognition stage.

DOI: 10.54254/2755-2721/121/2025.19493

During the following decades, in 1996, Bell South company developed the world's first Voice Activated Portal (VAL), an interactive voice recognition system for dial-in callers. It is designed to provide services over the phone, similar to today's automated voice menus. It provided information based on the user's voice input over the phone, setting a precedent for subsequent voice-activated menus, although these systems often frustrated users with inaccuracies for the next 15 years. By 2001, speech recognition technology had reached an accuracy rate of nearly 80%. However, the technology stagnated until Google launched the Google Voice Search application, which advanced speech recognition technology by analyzing large amounts of data to match user queries with actual speech examples using the massive computing power of data centers. The Siri project began as research at Stanford University's Center for Computing and Natural Language in 2003, aiming to create a virtual personal assistant that used natural language processing and machine learning to perform tasks[5]. In February 2010, Siri was launched as an iPhone app, helping users complete tasks such as sending text messages, booking restaurants, and searching online. Siri subsequently became a massive success on the iPhone 4S, being promoted as a personal assistant that could help users perform daily tasks through voice and gradually adapt based on the user's interests and preferences. As one of the first speech recognition systems to engage in conversation with users, Siri was also one of the early artificial intelligence virtual assistants.

In 2017, China company Xiaomi introduced a voice-controlled artificial intelligence interaction engine. Xiao Ai is Xiaomi's artificial intelligence interaction engine, widely used in Xiaomi mobile phones, speakers, TVs, computers, and other devices[6]. It supports 77 categories and over 4,000 smart devices and controls home appliances through voice commands. Xiao Ai has more than 1,400 skills, covering areas such as content, tools, and interaction. It is used in multiple scenarios, such as personal travel, smart homes, intelligent wearables, and bright offices., and it has become a ubiquitous intelligent assistant in users' lives. Xiaomi's acoustic technology team conducts research in call noise reduction, microphone arrays, and intelligent perception to support the acoustic algorithm needs of Xiaomi's various product lines. At the same time, Xiaomi's voice technology provides it with various voice understanding and generation technologies such as speech recognition, voiceprint recognition, and emotion recognition.

3. Traditional speech recognition applications

The traditional speech recognition learning method uses a statistical model and hand-designed feature extraction algorithm. The process typically began with the extraction of acoustic features from speech signals, such as Mel-frequency cepstral coefficients (MFCCs) and linear prediction coding (LPC), which effectively captured the speech’s acoustic properties. These methods excelled in simple and domain-specific tasks but struggled with complex and diverse speech inputs due to their dependence on manually designed features and limited context awareness. With the increase in data volume and speech recognition needs, deep learning-based technologies are gradually replacing traditional learning.

Automatic continuous speech recognition has many potential applications, including command and control, dictation, transcription of recordings, searching audio documents, and interactive spoken dialogue. At their core, all speech recognition systems rely on a set of statistical models used to represent the various acoustic features of language. Since speech signals have a time series structure and can be encoded as a sequence of spectral vectors spanning the audio frequency range, hidden Markov models are a natural choice for building these models. HMMs are central to modern speech recognition systems, and while their basic framework has not changed significantly in the past decade or so, the detailed modeling techniques developed around them have become increasingly sophisticated. The basic principle of HMMs is to create stochastic models from known utterances and compare unknown utterances with the probabilities generated by these models[1][7]. As a doubly

DOI: 10.54254/2755-2721/121/2025.19493

stochastic model suitable for non-stationary signals, HMMs allow other stochastic models to be plugged into the system to incorporate information from multiple hierarchical knowledge sources. HMMs are based on an enhanced version of the Markov chain model, representing the state probabilities of a sequence of random variables[8]. An essential assumption of Markov chains is that if one wants to predict future states in a sequence, the current state is the only one that matters, and the influence of previous states on the future is only propagated through the current state. This assumption is similar to weather forecasting: if you want to predict tomorrow's weather, you only need to refer to today's weather and do not need to look at yesterday's weather. Markov chains are suitable for calculating the probability of a series of observable events, while HMM extends this concept to those hidden events or situations that are not directly observable. For example, in natural language processing, part-of-speech tags are inferred from observable sequences of words rather than being directly observed. Scientists call these tags that cannot be directly observed "hidden events" and use HMM to analyze the relationship between these hidden events and observable events. HMM allows scientists to see visible events in the input and hidden events as causal factors in the probabilistic model. With the successive expansion of technology, HMM has reached a considerable level of complexity, promoting steady progress in speech recognition.

The Gaussian mixture model is an extension of a single Gaussian probability density function, which accurately quantifies the distribution of variables through multiple Gaussian distributions and standard distribution curves[9]. GMM decomposes the distribution of variables into multiple models based on Gaussian probability density functions and usually uses the expectation maximization algorithm for parameter estimation. It is a commonly used clustering algorithm that assumes that data points are generated by a mixture of multiple Gaussian distributions, each representing a different cluster or group in the data. GMM is precisely fit for scenarios where the underlying data distribution is unknown and can be represented as a combination of multiple Gaussian distributions. Mathematically, the probability density function of GMM is the weighted sum of K Gaussian components, where its mean vector and covariance matrix define each component. In machine learning, GMM is often used for clustering tasks, especially for data with elliptical clustering structures. In addition, GMM is also used in image processing, such as background subtraction and image segmentation, where different areas of the image can be modeled as different Gaussian distributions.

4. Modern Deep Learning Algorithms

In recent years, the field of speech processing has undergone tremendous changes due to the introduction of deep learning technology. Architectures such as deep neural networks, convolutional neural networks, and recurrent neural networks have greatly improved the effects of speech recognition, speaker recognition, and speech synthesis. Deep learning replaces manual feature engineering by automatically extracting meaningful features from speech signals, especially when dealing with noise, accents, and dialects[10]. Its powerful functions improve the adaptability and robustness of the system. Compared to traditional hidden Markov models, deep learning has markedly improved speech recognition accuracy. New technologies such as attention mechanisms and transformers have further improved the processing system's performance, enabling it to cope with long distances. Dependencies and complex patterns drive the progress of diverse speech-processing applications.

Convolutional neural networks are an essential subset of machine learning and the core of deep learning algorithms. They are usually composed of input, hidden, and output layers[11]. Among the various neural network types, feed forward networks are standard, while RNNs are prevalent in natural language processing and speech recognition. CNNs are extensively used in image classification and computer vision tasks. The advent of CNNs revolutionized image recognition by

DOI: 10.54254/2755-2721/121/2025.19493

leveraging matrix multiplication to identify image patterns, obviating the need for manual feature extraction. The convolution layer is the core part of CNN. Through a multi-layer structure, the network gradually shifts from simple features such as color and edges to identifying complex shapes and target objects. As the complexity of each layer increases, CNN can gradually parse larger structures and details in the image and finally accurately identify the target object. As a vital tool in computer technology, CNNs are applied across various domains, including autonomous driving and medical imaging, providing robust visual perception capabilities for artificial intelligence applications.

RNNs are deep learning neural networks designed for sequence or time series data processing, widely used in tasks like natural language processing, speech recognition, and image caption generation. Unlike traditional feed forward neural networks, RNNs can take advantage of the "memory" feature by retaining previous input information and affecting subsequent input and output[10]. This means that the output of RNN depends not only on the current input but also on the previous elements in the sequence. However, RNN also has some limitations, such as gradient vanishing and gradient explosion problems, which may cause the model to be unable to train and update parameters effectively. When the gradient disappears, the amplitude of the weight update gradually decreases until it can no longer learn, while the gradient explosion causes the weight to grow too large, eventually making the model invalid. To solve these problems, researchers introduced the extended short-term memory network. LSTM can control which information needs to be retained or forgotten by introducing three gating mechanisms in the hidden layer: input gate, output gate, and forget gate, thereby overcoming the problem that traditional RNN models cannot handle long-distance dependencies. For example, when processing language tasks, LSTM can remember critical contexts mentioned several sentences ago to help predict the content of the current sentence, while traditional RNNs may ignore this information because the context distance is too long. In addition, the gated recurrent unit is a simplified version of LSTM that reduces complexity by using reset gates and update gates while excellently processing long time series tasks. GRU effectively solves the short-term memory problem of RNN by controlling the flow of information. These advances allow RNNs, LSTMs, and GRUs to perform efficiently while dealing with complex time series tasks.

Semi-supervised learning is a machine learning technique that combines supervised and unsupervised learning and aims to reduce reliance on labeled data, especially in situations where labeling is costly or labeled samples are insufficient, such as in medical diagnosis. SSL uses a small amount of labeled data and a large amount of unlabeled data to train the model, solving the high cost and time problems of supervised learning that rely on a large amount of manually labeled data, as well as the shortcomings of unsupervised learning with limited application range and low accuracy[12]. By training an initial model on partially labeled data, SSL can be iteratively applied to unlabeled data, significantly improving the model's generalization capabilities[13]. Co-training is a technique in SSL that trains two classifiers based on two independent views of the data that provide additional information and are independent of each other, allowing efficient predictions even with small amounts of labeled data. Co-training performs well in tasks such as web page classification, for example, by classifying a web page's text and link anchor views. Semi-supervised learning shares a similar goal with transfer learning, namely reducing the labeling burden, but they are often studied separately, with SSL usually trained from scratch, while transfer learning is initialized using a pre-trained model. In automatic speech recognition (ASR) systems, generating narrow output distributions is a common challenge, closely related to the connectionist temporal classification (CTC) method. CTC is a powerful sequence learning tool that aligns sequences without explicit alignment, handling variable-length sequences and consecutive identical characters by introducing blank symbols[14,15]. CTC has significantly impacted fields like speech recognition, text recognition, and video segmentation. However, there are still several problems with model overconfidence, which has

DOI: 10.54254/2755-2721/121/2025.19493

enlightened scientists to continue to develop various regularization techniques to overcome the problem of overconfidence. Label smoothing regularization is an effective method that maximizes the relative entropy between the model prediction and the uniform distribution using soft targets instead of challenging targets. In addition, confidence regularization is also related to the maximum entropy principle, which advocates that maximizing uncertainty can better describe the current state under given constraints. EnCTC is a CTC regularization method based on maximum conditional entropy, which aims to prevent the entropy of feasible paths from decreasing too quickly during training. Thereby mitigating the impact of CTC converging to a single path and encouraging the model to explore more during training. These regularization strategies help improve the performance and reliability of ASR systems.

In the future, speech recognition systems may incorporate advanced biometric technologies, understand speech content, identify the speaker, and significantly improve data security and accuracy. This unique sound signature promises to be a new means of two-factor authentication. Recently, speech recognition technology has made significant progress in noise reduction, speaker differentiation, and natural language processing, with a dramatic improvement in the system's accuracy and efficiency. Future artificial intelligence voice assistants will better understand the context and respond more accurately based on the user's geographical location, time, and past interaction history. However, these technological advances still need to address some challenges, such as misunderstandings caused by background noise, confusion about homophones, handling of speech impediments, and privacy issues related to voice data recording and processing.

5. Conclusion

Speech recognition technology has evolved significantly since the 1950s and has become integral to modern applications. It enables natural, hands-free interaction by converting spoken words into text, improving people's communication efficiency and device control convenience. With the advancement of artificial intelligence and machine learning, speech recognition accuracy has dramatically improved, especially in noise reduction, speaker distinction, and natural language processing. Users can now interact with virtual assistants via voice to perform tasks such as checking the weather or adjusting settings, as well as setting alarms and managing smart home devices. Nevertheless, speech recognition technology grapples with numerous challenges, notably the complexity of human language and pronunciation variations, which complicates accurate transcription. The most common techniques, such as Hidden Markov Models and Gaussian Mixture Models, have limitations in handling different accents and dialects. Machine learning techniques such as deep neural networks, convolutional neural networks, and recurrent neural networks can automatically extract meaningful features, improving the robustness and adaptability of the system. Overall, speech recognition technology and algorithms are still evolving and probably still for the next couple of decades. In the future, it will be more intelligent and humanized to provide better user experience.

References

[1] *Burchi, Maxime, and Vielzeuf, Valentin. Efficient Conformer: Progressive Downsampling and Grouped Attention for Automatic Speech Recognition. 2021.*
[2] *"A Brief History of Speech Recognition." Sonix. Accessed 5 Sept. 2024.*
[3] *"Speech Recognition." IBM. Accessed 7 Sept. 2024.*
[4] *The Harpy Speech Recognition System, Stanford University. Accessed 9 Sept. 2024.*
[5] *Media, OpenSystems. "The Invention of Apple's Siri and Other Virtual Assistants." Embedded Computing Design. Accessed 9 Sept. 2024.*
[6] *"AI xiaoai." Mi.com, 2024, Accessed 12 Sept. 2024.*
[7] *Juang, B. H., & Rabiner, L. R. (1991). Hidden Markov Models for Speech Recognition. Technometrics, 33(3), 251–272. Accessed 14 Sept. 2024.*

DOI: 10.54254/2755-2721/121/2025.19493

[8] Gales, Mark, and Steve Young. The Application of Hidden Markov Models In Speech Recognition. Hanover, Ma, Now Publishers, Cop, 2008. Accessed 14 Sept. 2024.
[9] "What Is: Gaussian Mixture Model." LEARN STATISTICS EASILY, 2024. Accessed 14 Sept. 2024.
[10] Ambuj Mehrish, Navonil Majumder, Rishabh Bharadwaj, Rada Mihalcea, Soujanya Poria, A review of deep learning techniques for speech processing, Information Fusion, Volume 99, 2023, 101869, ISSN 1566-2535.
[11] "What Is: Gaussian Mixture Model." LEARN STATISTICS EASILY, 2024. Accessed 14 Sept. 2024.
[12] IBM. "What Are Convolutional Neural Networks? | IBM." IBM, 2024. Accessed 15 Sept. 2024.
[13] Altexsoft. "Semi-Supervised Learning, Explained with Examples." AltexSoft, 18 Mar. 2022. Accessed 15 Sept. 2024.
[14] Daniel, Jurafsky, and James Martin. Speech and Language Processing. 7 Jan. 2023. Accessed 14 Sept. 2024.
[15] "AdaMER-CTC: Connectionist Temporal Classification with Adaptive Maximum Entropy Regularization for Automatic Speech Recognition." Arxiv.org, 2024, Accessed 17 Sept. 2024.