Innovative Deep Learning Models for Speech Recognition

Meghana Udiga

Master's in Business Analytics

Department of Management Information Systems

Abstract

Deep learning models have been instrumental in the major developments in voice recognition technology over the past few years, particularly in terms of accuracy. This technology makes it possible for people and machines to communicate naturally and intuitively, and it has found use in security, transportation, and the healthcare industries. This study compares the effectiveness of contemporary voice assistants like Google Home, Siri, and Alexa with various deep learning models for speech recognition, including DNN, DBN, and RNN. The essay also looks at the field's potential advancements and limitations. Speech recognition technology has made great strides, and more study and development in this area is anticipated to result in major enhancements in functionality, accuracy, and dependability, making it a promising area for future innovation.

1. Introduction

It is impossible to exaggerate the value of speech recognition in the area of human-computer interaction. It makes sense that the next technical advancement in human-computer interaction (HCI) will be natural language voice recognition as speech is the most efficient and natural mode of human communication. Using computer programs and algorithms, speech recognition involves turning a vocal signal into a string of words. The development of methods and systems for speech input to machines is the aim of speech recognition. Automatic speech recognition is now widely used in jobs that call for human-machine interfaces, like automatic call processing, thanks to recent advances in statistical modelling of speech.

Since the 1960s, voice recognition has been the subject of computer science research. Early attempts were crude, and the first systems that could understand speech didn't appear until the 1980s. These early systems' capabilities and power, however, were constrained. Speaking is the primary mode of human communication; therefore, it makes sense that people would want speech interfaces with computers that can speak and understand their native tongues. The process of creating a string of words that most closely resembles a particular audio signal is known as machine speech recognition. Speech recognition has a wide range of known uses, including virtual reality, multimedia searches, auto-attendants, travel planning and reservations, translators, NLU, and many more.

With a particular application in mind, we will concentrate in this research paper on cutting-edge deep learning models for speech recognition. We will examine the most recent methods and algorithms in use, their efficacy, as well as the difficulties and constraints in this area. We'll also talk about deep learning's commercial uses and any prospective advancements in the application we've chosen. Additionally, the existing drawbacks of deep learning models in this area will be evaluated, along with possible remedies. This study seeks to give a thorough overview of recent advancements in speech recognition, especially with regard to deep learning models, and their potential for use in the future.

2. Literature Review

With substantial developments in deep learning techniques in recent years, speech recognition has advanced significantly from its infancy in the 1960s. We shall summarize the state-of-the-art deep learning models and algorithms used in voice recognition today, their efficacy, and the difficulties and constraints in this field in this review of the literature.

The two main deep learning models for speech recognition are recurrent neural networks (RNNs) and convolutional neural networks (CNNs). Acoustic modelling has made use of CNNs, which develop feature maps by extracting characteristics from audio signals. These feature maps are then fed into fully connected neural networks, which provide probability distributions over speech units. In low-resource languages and speaker-independent recognition challenges, this method has shown to be successful.

RNNs, however, are employed to simulate the temporal dependencies present in voice data. The RNN architectures known as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) have been applied to speech recognition. While GRUs have been discovered to be more effective and quicker to train, LSTMs have demonstrated encouraging results in speaker-independent identification tasks.

In the acoustic and linguistic modelling stages of speech recognition, Deep Neural Networks (DNNs) have also been employed. To categorize speech units and produce a probability distribution for the following unit in the sequence, DNNs are trained. In tasks requiring continuous voice recognition with a wide vocabulary, this method has proven successful.

The quality and quantity of the training data have a significant impact on how well these deep learning models and algorithms perform. A significant problem in this area, particularly for low-resource languages, is the dearth of labelled data. To solve this problem, domain adaptation and transfer learning strategies have been suggested, where models trained on a big dataset are improved on a smaller dataset.

Additionally, a hurdle with voice recognition systems is their resilience to outside noise and speaker fluctuation. To solve this problem, data augmentation methods such introducing background noise and speaker variety have been suggested.

Many different businesses, including healthcare, transportation, and security, use speech recognition in unique ways. Speech recognition is used in healthcare for telemedicine, clinical documentation, and patient interaction. Speech recognition is utilized in the transportation industry for driver assistance, navigation, and in-car entertainment. Speech recognition is used in security for forensic investigations, access control, and speaker identification.

The use of deep learning models in end-to-end speech recognition, where the acoustic and linguistic modelling stages are merged into a single model, is one of the upcoming developments in voice recognition. Recent studies have demonstrated encouraging outcomes for this strategy. Additionally, it is anticipated that the robustness and generalization of speech recognition systems would be enhanced by the application of unsupervised learning approaches like transfer learning. The technology of speech recognition has substantially evolved thanks to deep learning models and algorithms, but there are still issues and limitations, particularly with low-resource languages, background noise, and speaker unpredictability. The use of voice recognition in several industries demonstrates how effective and efficient it can be when used with people. The performance and resilience of will likely be significantly improved by improvements in end-to-end speech recognition and unsupervised learning approaches in the future.

3. Industry Applications

There are numerous possible uses for speech recognition technologies in numerous sectors. Here are some illustrations.

Healthcare: Healthcare practitioners can easily and reliably capture patient information by using speech recognition to transcribe medical dictation. Additionally, it can be used to automate some administrative operations, such making appointments and purchasing supplies.

Transportation: Speech recognition can be used in the transportation sector to increase safety by enabling drivers to make phone calls or switch radio stations while keeping their hands on the wheel and their eyes on the road.

Security: Based on a person's voiceprint, speech recognition can be utilized in security applications like access control systems to confirm that person's identification.

Customer service: By allowing customers to engage with automated systems using their voice rather than having to negotiate difficult menus and options, speech recognition can be utilized to improve the customer experience in the field of customer service.

Finance: Automating financial processes, including stock trading, and giving consumers real-time financial advice are both possible with speech recognition.

Manufacturing: Speech recognition can be used to increase productivity and worker security in the manufacturing industry. Without taking their hands off their tools, workers can use voice commands to manage equipment, check inventory, and report safety incidents.

Retail: In the retail industry, speech recognition can be used to enhance customer service by enabling customers to place orders, ask questions, and offer comments using their voice. By allowing staff to update and check inventory using voice commands, it can also be utilized to improve inventory management.

Education: The use of speech recognition technology can increase the accessibility of learning environments for students with impairments. To make lectures and class discussions simpler to follow for students who have hearing difficulties, it can be utilized to automatically transcribe them.

Speech recognition can be utilized in law enforcement to automate several administrative activities, including processing warrants and filling out incident reports. By enabling cops to use voice commands to operate their in-car technology, such radios, and sirens, it can also be used to increase officer safety.

Future developments in technology are likely to bring in even more creative applications.

4. NEUTRAL NETWORKS

A. Overview of Neural Networks

Mathematical models called neural networks imitate how the human brain functions. They compute likely solutions based on various probabilities by using estimate functions, or neurons. To build a network of neurons that simulate a specific function, these neurons are then used in another function with varied probabilities. This makes it possible to produce results that are better and more accurate. Machine learning uses neural networks to educate computers to carry out tasks based on training examples. A machine, for instance, can be taught the sound of the word "Hello" by listening to recordings of individuals pronouncing it in a variety of accents, pitches, and background noise levels.

B. Deep Neural Networks

Unsupervised feature learning or representation learning is a key component of the emerging machine learning field known as deep learning. The neural networks employed in the gaming industry, which utilize graphics processing units (GPUs) to carry out complicated computations, gave rise to deep neural networks (DNNs). DNNs have replaced Gaussian mixture models as the standard technique for speech recognition. They enable effective and natural discriminative training by estimating probability of speech feature segments. With the decline in costs and rise in computer capacity, brute force training on huge datasets has become more practical. Deep Belief Networks (DBNs) are a subclass of DNN that are taught unsupervised one at a time using a stack of constrained Boltzmann machine layers.

C. Deep Belief Network

DBNs are neural networks made up of a stack of restricted Boltzmann machine (RBM) layers that are trained one at a time, unsupervised, to produce progressively more abstract representations of the inputs in succeeding layers. A DLN is essentially a stack of RBMs, each of which has two layers. Unsupervised pre-training is a greedy learning technique used in unsupervised training. After the unsupervised training is finished, the supervised training starts, which modifies the weights produced using gradient descent learning to enhance the performance of DBN.

D. Recurrent Neural Networks

RNNs, or recurrent neural networks, are neural networks that could store their present state each time they receive input. As a result, RNNs can effectively model sequential data and generate predictions based on historical inputs. Applications involving sequential data, such as speech recognition and natural language processing, can benefit from the use of RNNs. When training RNNs on lengthy data sequences, the vanishing gradient problem might occur. Long Short-Term Memory (LSTM) is a form of RNN that can prevent this.

5. Current Implementations

5.1 Google Home

A home assistant called Google Home makes use of automatic speech recognition technology. It is one of the top systems on the market that processes user requests using neural network models. The system's ability to comprehend user demands, particularly in noisy environments, is enhanced using multichannel processing, acoustic modelling, and Grid-LSTMs to simulate frequency fluctuations.

With the use of its own unique data and a grid with long short-term memory, Google Home can effectively model frequency changes. The device needs an internet connection because the available technology is insufficient to calculate the data on its own and instead relies on cloud computing.

The public has learned to trust Google Home, and the company's neural networks and algorithms are now strong enough to respond to user queries in an efficient and accurate manner. It has become a popular technology for many households thanks to its capacity for understanding natural language and carrying out a variety of tasks.

5.2 Siri

Apple developed the virtual assistant Siri, which debuted in 2011. It enables consumers to communicate with their Apple gadgets using voice commands in natural language. Users can activate the assistant via the "Hey Siri" feature without having to touch their smartphone. Siri recognizes the user's voice and understands their orders using a deep neural network. The system recognizes the user's voice and determines if they have issued a command using a two-step method. A simple voice recognizer that is constantly operating and listening for the word "Hey Siri" is the first stage. In order to ascertain whether the user truly intended to voice-activate Siri, the second step entails a temporal integration procedure that computes a confidence score. When the score reaches a certain level, the device will wait for a command. Siri is capable of a wide range of operations, such as placing calls, sending texts, creating reminders, and playing music.

5.3 Alexa

Amazon developed the voice-activated virtual assistant Alexa, which functions similarly to Google Home and Siri. It employs internet connectivity and cloud-based deep neural networks to process and compute data retrieved from the system at the user's house. For each voice query, Alexa employs a two-step, scalable, and effective neural shortlisting-reranking approach to determine which skill is the most pertinent. Since its launch in 2014, Alexa has been the best-selling voice-powered AI gadget in the US, with Amazon reportedly accounting for around 70% of all unit sales. As a result of Alexa's popularity, many now find it difficult to live without her, and some even incorporate her into their daily routines. Despite some setbacks along the way, such as kids ordering items through Alexa, the big names in technology have developed solutions like parental controls.

6. Performance of the systems

Depending on the precise tasks or commands the user issues to Google Home, Alexa, or Siri, their performance may change. All three systems have, however, generally displayed excellent competence in their respective fields.

Google Home is renowned for its capacity to provide precise answers to queries and carry out general knowledge and search-related tasks. It can do a lot of things, like set alarms and reminders and manage smart home appliances. Additionally, Google Home features a natural language processing engine that can comprehend difficult questions and provide answers.

In contrast, Alexa is renowned for its enormous collection of skills, which are essentially external apps that can be utilized with the gadget. With the use of these abilities, Alexa can carry out a variety of tasks, including ordering meals, playing music, operating home appliances, and even booking an Uber. Additionally, Alexa has superb speech recognition technology that enables it to effectively understand and carry out requests.

Apple's iOS devices include Siri, which is well-known for its personal assistant functions. It is capable of carrying out operations including placing calls, sending messages, creating reminders, and arranging appointments. Apple apps like Apple Music and Apple Maps can also be integrated with Siri. The ability of Siri to operate seamlessly with other Apple products and services depends on its ecosystem integration.

Overall, all three systems operate and function quite well, however based on the requirements and preferences of the user, different characteristics and advantages of each system may apply.

7. Future Developments, Limitations and Solutions

Natural language processing, machine learning, and speech recognition might all see advancements in the next years for voice-based AI systems like Siri, Alexa, and Google Home. These advancements may enable more precise and responsive user interactions as well as the capacity to comprehend and reply to more intricate requests and enquiries.

These methods do have drawbacks though, and these must be addressed. Their dependency on internet access, which might have problems with speed and dependability, is one constraint. Their potential to violate user privacy due to their constant listening for voice instructions and potential collection and use of that data for other purposes is another drawback.

These restrictions might be overcome by enhancing offline functionality, enforcing tougher user controls and privacy regulations, and introducing more sophisticated security mechanisms. Further study and development may also result in more complex algorithms and models that are less dependent on internet access and offer better user privacy protection.

Overall, voice-based AI systems have advanced significantly over the past few years and have been incorporated more and more into our daily lives. To address these systems' shortcomings and difficulties, there is still opportunity for advancement and a need for ongoing innovation and development.

8. Conclusion

In conclusion, deep learning and the creation of neural networks have contributed to the rapid advancement of voice recognition technologies in recent years. These innovations have made speech recognition more precise and reliable, which has led to a wide range of new applications in numerous industries. Speech recognition is being utilized to automate operations, increase safety, and improve the customer experience in a variety of industries, including healthcare, transportation, security, and customer service. Speech recognition will probably be used in increasingly more advanced and potent ways in the years to come as neural networks and other machine learning technologies continue to advance. The potential for speech recognition technology to change how we live, and work is thrilling.

9. References

[1] R.Klevansand R.Rodman, "Voice Recognition, Artech House, Boston, London 1997.

[2] Gerhard Rogoll,Maximum Mutual Information Neural Networks for hybrid connectionist-HMM speech Recognition systems ,IEEE Transaction on Audio, speech and Language Processing Vol.2 ,No.1,Part II,Jan.1994.

[3] Antonio M. Peinado et.al, discriminative codebook design using Multiple Vector quatization in HMM based speech recognizers,IEEE Transaction on Audio,Speech and language Processing Vol.4 No.2 March.1996

[4] Nam Soo kim et.al,On estimating robust Probability Distribution in HMM in HMM based Speech Recognition ,IEEE Transaction on Audio, Speech and Language Processing Vol.3,No.4 ,July 1995.

[5] Jean Francois, Automatic word Recognition Based on Second Order hidden Markov Models.IEEE Transaction on Audio, Speech and Language ProcessingVol.5, No.1, Jan.1997.

[6] Samudravijaya K. Speech and Speaker recognition tutorial TIFR Mumbai 400005.

[7] Kevin Brady, Michael Brandstein, Thomas Quatieri, Bob Dunn "An Evaluation Of Audio-Visual person Recognition on the XM2VTS corpus using the Lausanne protocol" MIT Lincoln Laboratory, 244 Wood St., Lexington MA

[8] W. M. Campbell_, D. E. Sturim W. Shen D. A. Reynolds_,
J. Navr´atily "The MIT- LL/IBM Speaker recognition System using High performance reduced Complexity recognition" MIT Lincoln Laboratory IBM 2006.

[9] M.J.F.Gales and S.J young, Parallel Model combination for Speech Recognition in Noise technical Report, CUED/FINEFENG/TRI135, 1993.

[10] A.P.Varga and R.K.Moore, "Hidden Markov Model Decomposition of Speech and Noise, Proc.ICASSp, pp.845- 848, 1990.

[11] M.Weintraub et.al, linguistic constraints in hidden markov Model based speech recognition, Proc.ICASSP, pp.699-702, 1989.

[12] S.katagiri, Speech Pattern recognition using Neural Networks.

[13] L.R.Rabiner and B.H.jaung ," Fundamentles of Speech Recognition Prentice-Hall, Englewood Cliff, New Jersy, 1993

[14] Zahi N.Karam,William M.Campbell "A new Kernel for SVM MIIR based Speaker recognition "MIT Lincoln Laboratory, Lexington, MA, USA.

[15] Asghar .Taheri ,Mohammad Reza Trihi et.al,Fuzzy Hidden Markov Models for speech recognition on based FEM Algorithm, Transaction on engineering Computing and Technology V4 February 2005,IISN,1305-5313

[16] GIN-DER WU AND YING LEI " A Register Array based Low power FFT Processor for speech recognition" Department of Electrical engineering national Chi Nan university Puli ,545 Taiwan

Technology, Kharagpur Kharagpur-721302 West Bengal,India. .

[17] Kenneth Thomas Schutte "Parts-based Models and Local Features for Automatic Speech Recognition" B.S., University of Illinois at Urbana-Champaign (2001) S.M., Massachusetts Institute of Technology (2003).

[18] Zaidi Razak, Noor Jamaliah Ibrahim, Emran Mohd Tamil, Mohd Yamani Idna Idris "Quarnic Verse recitation feature extraction using Mel-Frequency Cepstral Coefficient(MFCC)" Department of Al-Quran & Al-Hadith, AcademyOf Islamic Studies, University of Malaya .

[19] Samudravijay K "Speech and Speaker recognition report" source: http://cs.jounsuu.fi/pages/tkinnu/reaserch/index.html

Viewed on 23 Feb. 2010.

[20] Sannella, M Speaker recognition Project Report report" From http://cs.joensuu.fi/pages/tkinnu/research/index.html Viewed 23 Feb. 2010.

[21] IBM (2010) online IBM Research

Source:-
http://www.research.ibm.com/Viewed 12 Jan 2010.

[22] Nicolás Morales1, John H. L. Hansen2 and Doorstep T. Toledano1 "MFCC Compensation for improved recognition filtered and band limited speech" Center for Spoken Language Research, University of Colorado at Boulder,
Boulder (CO), USA

[23] M.A.Anusuya , S.K.Katti "Speech Recognition by Machine: A Review" International journal of computer science and Information Security 2009.

[24] Goutam Saha, Ulla S. Yadhunandan " Modifield Mel- Frequency Cepstral coefficient Department of Electronics and Electrical communication Engineering India Institute of

[25] P.satyanarayana "short segment analysis of speech for enhancement" institute of IIT Madras feb.2009

[26] David, E., and Selfridge, O., Eyes and ears for computers, Proc.IRE 50:1093.

[27] SadokiFuruki,Tomohisa Ichiba et.al,Cluster-based Modeling for Ubiquitous Speech Recognition, Department of Computer Science Tokyo Institute of Technology Interspeech 2005.

[28] Spector, Simon Kinga and Joe Frankel, Recognition ,Speech production knowledge in automatic speech recognition ,
Journal of Acoustic Society of America,2006

[29] M.A Zissman,"Predicting,diagonosing and improving automatic Language identification performance" ,Proc.Eurospeech97,Sept.1997 vol.1,pp.51-54 1989.

[30] Y.Yan and E.Bernard ,"An apporch to automatic language identification basedon language depandant phone recognition
",ICASSP'95,vol.5,May.1995 p.3511