State-of-the-art Deep Learning Models for Speech Recognition

Venu Dodda

Master's in Business Analytics

Department of Management Information Systems

Abstract

Speech recognition technology has advanced significantly in recent years, with deep learning models playing a critical role in boosting accuracy. This technology, which allows for natural and intuitive interactions between humans and robots, has implications in healthcare, transportation, and security. This research investigates the various deep learning models used in speech recognition, such as DNN, DBN, and RNN, and assesses the performance of existing implementations such as Google Home, Siri, and Alexa. The article also looks at their limitations and the field's potential advances. Speech recognition technology breakthroughs have immense promise, and continued research and development in this sector is likely to result in major increases in accuracy, dependability, and functionality, making it a promising area for future innovation.

1. Introduction

In the subject of human-computer interaction, the importance of voice recognition cannot be emphasized. Speech is the most natural and efficient method of human communication; hence natural language speech recognition should be the next technological advancement in HCI. The process of turning a speech signal into a sequence of words using algorithms implemented as computer programs is known as speech recognition. The purpose of voice recognition is to create techniques and systems for machine speech input. Advances in statistical modelling of speech have resulted in the widespread use of automatic speech recognition in tasks requiring human-machine interfaces, such as automatic call processing.

Since the 1960s, computer scientists have been working on voice recognition. Early attempts were crude, and it wasn't until the 1980s that the first speech-deciphering systems appeared. The scope and capability of these early systems, however, were restricted. People naturally expect voice interfaces with computers that can speak and recognize speech in their native language because spoken language dominates human communication. The process of creating a sequence of words that best matches the provided speech signal is known as machine speech recognition. Virtual reality, multimedia searches, auto-attendants, travel information and reservations, translators, natural language understanding, and many other applications are known to use voice recognition.

In this study, we will concentrate on cutting-edge deep learning models for speech recognition, with a specific application in mind. We will go over the most recent methodologies and algorithms, as well as their usefulness, problems, and limitations in this sector. We will also talk about industry uses of deep learning and potential future developments in our chosen application. The existing constraints of deep learning models in this field, as well as potential remedies to these restrictions, will be investigated. This research study seeks to present a complete overview of the most

recent advances in speech recognition, notably deep learning models, and their potential for future applications.

2. Literature Review

Since its inception in the 1960s, speech recognition has come a long way, with notable advances in deep learning techniques in recent years. We will summarize the present state-of-the-art deep learning models and algorithms used in voice recognition, their effectiveness, and the challenges and limitations in this field in this literature review.

The two basic deep learning models used in speech recognition are Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs have been used in acoustic modelling to extract features from audio signals and build a feature map, which is then sent through a fully connected neural network to generate a probability distribution over the speech units. This method has been shown to be effective in low-resource languages as well as speaker-independent recognition challenges.

In contrast, RNNs are employed to model temporal dependencies in voice signals. Popular RNN designs used in voice recognition include Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). In speaker-independent recognition tasks, LSTMs have showed promising results, whereas GRUs have been shown to be more efficient and faster to train.

Deep Neural Networks (DNNs) have also been employed in voice recognition, especially during the acoustic and language modelling stages. DNNs are trained to recognize speech units and create a probability distribution for the following unit in a sequence. This method has been demonstrated to be effective in large vocabulary continuous voice recognition tests.

The performance of these deep learning models and algorithms is strongly dependent on the quality and quantity of training data. The absence of labelled data is a significant barrier in this subject, particularly for low-resource languages. Domain adaptation and transfer learning strategies have been developed to overcome this issue, in which models trained on a big dataset are fine-tuned on a smaller dataset.

Furthermore, the resilience of speech recognition systems against external noise and speaker unpredictability is also an issue. To solve this issue, data augmentation approaches such as introducing background noise and speaker variation have been proposed.

Speech recognition is used in a variety of industries, including healthcare, transportation, and security. Speech recognition is utilized in clinical recordkeeping, telemedicine, and patient interaction in healthcare. Speech recognition is used in transportation for in-car infotainment, navigation, and driver assistance. Speech recognition is used in security for speaker identification, access control, and forensic investigations.

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 speech recognition. Recent research has yielded encouraging outcomes using this strategy. Furthermore, the application of unsupervised and transfer learning approaches is likely to improve the resilience and generalization of voice recognition systems.

In conclusion, deep learning models and algorithms have advanced voice recognition technology tremendously, but obstacles and limitations remain, notably in low-resource languages, environmental noise, and speaker

variability. The use of voice recognition in various industries demonstrates its potential to increase human-machine interaction and efficiency. Future advances in end-to-end speech recognition and unsupervised learning approaches are projected to improve the performance and reliability of the system.

3. Industry Applications

Speech recognition technology has numerous potential uses in a variety of industries. Here are a couple of such instances.

Healthcare: Speech recognition can be used to transcribe medical dictation, allowing healthcare practitioners to record patient information easily and accurately. It can also be used to automate administrative chores like appointment scheduling and procuring supplies.

Transportation: Speech recognition can be used to improve safety in the transportation business by allowing drivers to maintain their hands on the wheel and their eyes on the road while still performing duties such as making phone calls or changing radio stations.

Speech recognition can be used in security applications such as access control systems to authenticate an individual's identification based on their voiceprint.

Customer Service: Speech recognition can be used to improve customer service by allowing customers to engage with automated systems using their voice rather than having to navigate complex menus and options.

Finance: Speech recognition can be used to automate financial operations like stock trading and to give customers real-time financial advice.

Manufacturing: Speech recognition can be used to boost production and worker safety in the manufacturing industry. Workers, for example, can use voice commands to manage equipment, check inventory, and record safety incidents without taking their hands off their tools.

Retail: In retail, speech recognition can be used to improve customer service by allowing customers to place orders, inquire, and provide comments using their voice. It can also be used to improve inventory management by allowing employees to utilize voice commands to execute inventory checks and changes.

Education: Speech recognition technology can be used to improve accessibility for students with impairments in the classroom. It can, for example, be used to automatically transcribe lectures and classroom discussions, allowing students with hearing problems to follow along more easily.

Speech recognition can be utilized in law enforcement to automate administrative chores such as filling out incident reports and processing warrants. It can also be utilized to increase officer safety by allowing officers to control their in-car technology, including as radios and sirens, using voice commands.

As technology advances, we should expect to see even more imaginative applications in the future.

4. NEURAL NETWORKS

A. Neural Network Introduction

Neural networks are mathematical models that simulate how the human brain operates. They employ estimate functions known as neurons to generate likely replies based on various probabilities. These neurons are then employed with varied probabilities in another function to form a network of neurons that simulates a certain function. This enables the production of better and more precise products. In machine learning, neural networks

are used to train computers to do tasks based on training examples. A machine, for example, can learn what the word "Hello" sounds like by hearing it spoken in various accents, pitches, and degrees of background noise.

B. Deep Neural Networks

Deep learning is a relatively recent branch of machine learning that entails unsupervised feature learning or representation learning. Deep Neural Networks (DNNs) evolved from neural networks used in gaming, which use graphics processing units (GPUs) to execute sophisticated computations. DNNs have surpassed Gaussian mixture models as the standard technique for speech recognition. They can estimate the probability of speech feature segments and enable natural and efficient discriminative training. With the decrease in expenses and increase in computing capacity, brute force training on huge datasets has become more viable. Deep Belief Networks (DBNs) are a sort of DNN that is made up of a stack of constrained Boltzmann machine layers that are taught unsupervised one at a time.

C. Network of Deep Beliefs

DBNs are neural networks made up of a stack of restricted Boltzmann machine (RBM) layers that are trained unsupervised one at a time to generate progressively abstract representations of the inputs in subsequent levels. A DLN is essentially a stack of RBMs, with each machine having two layers. Unsupervised pre-training is a greedy learning approach used in unsupervised training. After the unsupervised training is finished, the supervised training begins, which uses gradient descent learning to change the generated weights to improve DBN performance.

D. Recurrent Neural Network (RNN)

Recurrent Neural Networks (RNNs) are neural networks that can save their state after each input. RNNs can thus be a powerful model for sequential data, making predictions based on past inputs. RNNs can be utilized in natural language processing, speech recognition, and other sequential data applications. extended Short-Term Memory (LSTM) RNNs can overcome the vanishing gradient problem that happens while training RNNs on extended data sequences.

5. Current Implementations

5.1. Google Home

Google Home is a smart home assistant that uses automatic speech recognition technologies. It is one of the market's leading systems for processing user requests using neural network models. To predict frequency changes, the system employs a multichannel processing technique paired with acoustic modelling and Grid-LSTMs, allowing it to better interpret user demands, particularly in noisy environments.

Google Home changes its model based on its own data, and its extremely large short-term memory enables it to efficiently represent frequency variations. The device requires an internet connection because present technology is insufficient to compute the data on its own and requires cloud computing to do it.

Google Home has become a credible source for the public, and its neural networks and algorithms have developed sophisticated enough to deliver accurate and effective solutions to user queries. It has become a popular technology in many families due to its capacity to interpret natural language and execute a variety of functions.

5.2. Siri

Siri is an Apple virtual assistant that was initially presented in 2011. It enables consumers to communicate with their Apple gadgets using natural language voice commands. The "Hey Siri" feature allows users to summon the assistant without physically touching their smartphone. Siri recognizes the user's voice and interprets their orders using a deep neural network. The technology recognizes the user's voice and determines if they have issued a command in two steps. The first stage entails always running a small speech recognizer and listening for the word "Hey Siri." The second stage entails a temporal integration method that computes a confidence score in order to evaluate whether the user intended to call Siri using their voice. If the score is sufficiently high, the device will wait for a command. Siri can do a range of things, such as make phone calls, send text messages, create reminders, and play music.

5.3. Alexa

Alexa is an Amazon voice-controlled virtual assistant that functions similarly to Google Home and Siri. It processes and computes data obtained from the system at the user's house using cloud-based deep neural networks, which require an internet connection to function. Alexa finds the best relevant skill for a given voice query using a two-step, scalable, and efficient neural shortlisting-reranking approach. Alexa has become the top seller in the US market for voice-powered AI devices since its debut in 2014, with Amazon accounting for over 70% of all unit sales. Alexa has proven to be a popular gift among users, who now find it difficult to live without it, with some even incorporating her into their daily routines. Despite some glitches along the way, such as youngsters ordering toys through Alexa, technological behemoths have devised solutions such as parental controls.

6. Performance of the systems

Google Home, Alexa, and Siri's performance can vary based on the tasks or commands supplied by the user. However, in general, all three systems have performed admirably in their respective fields.

Google Home is well-known for its ability to answer queries accurately and execute tasks linked to search and general knowledge. It boasts a wide range of functions, including the ability to control smart home devices and create reminders and alarms. Google Home also includes a natural language processing system capable of comprehending and responding to sophisticated requests.

Alexa, on the other hand, is recognized for its extensive skill library, which consists of third-party apps that may be utilized with the device. These abilities enable Alexa to do everything from play music and order food to operate house appliances and even book an Uber. Alexa also has an outstanding speech recognition system, which allows it to accurately interpret and respond to requests.

Siri, which is built into Apple's iOS devices, is well-known for its personal assistant features. It can make calls, send messages, set reminders, and schedule appointments, among other things. Siri can also work with other Apple apps like Apple Music and Apple Maps. Siri's strength is its connection with Apple's ecosystem, which allows it to interact with other Apple products and services smoothly.

Overall, all three systems provide outstanding performance and capabilities; however, the specific features and strengths of each system may differ based on the user's demands and preferences.

7. Future developments, Limitations and Solutions

Natural language processing, machine learning, and speech recognition could all improve in the

future for voice-based AI systems like Siri, Alexa, and Google Home. These enhancements may result in more accurate and responsive user interactions, as well as the ability to interpret and respond to more complicated requests and questions.

These systems, however, have limits that must be addressed. Their dependency on internet access is one drawback, which can cause speed and reliability concerns. Another disadvantage is that they have the potential to violate user privacy because they constantly listen for voice instructions and may collect and utilize that data for other purposes.

Improving offline capabilities, introducing stronger privacy policies and user controls, and incorporating more modern security measures could all be solutions to these restrictions. Furthermore, additional research and development could result in more advanced algorithms and models that are less reliant on internet access and safeguard user privacy better.

Overall, voice-based AI systems have advanced significantly in recent years and are increasingly interwoven into our daily lives. However, there is still potential for improvement and a need for ongoing innovation and research to address the constraints and challenges that these systems present.

8. Conclusion

Finally, thanks to the development of neural networks and deep learning techniques, voice recognition technology has evolved dramatically in recent years. These technologies have enabled more precise and dependable speech recognition, opening the door to a plethora of new applications in a variety of industries. Speech recognition is being utilized to automate operations, increase safety, and improve the customer experience in industries ranging from healthcare to transportation, security to customer service. With the continuous development of neural networks and other machine learning technologies, we may expect to see even more inventive and powerful voice recognition applications in the coming years. The promise of voice recognition technology to revolutionize the way we live and work is intriguing.

9. References

[1] Samudravijay K "Speech and Speaker recognition report" source: http://cs.jounsuu.fi/pages/tkinnu/reaserch/index.html
Viewed on 23 Feb. 2010.

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

[3] IBM (2010) online IBM Research
Source:- http://www.research.ibm.com/Viewed 12 Jan 2010.

[4] 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

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

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

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

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

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

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

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

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

[13] Tavel R.K.Moore,Twenty things we still don't know about speech proc.CRIM/FORWISS Workshop on Progress and Prospects of speech Research and Technology 1994.

[14] H.Sakoe and S.Chiba, Dynamic programming algorithm optimization for spoken word recognition ,IEEE Trans. Acoustics, Speech, Signal Proc.,ASSP-26(1).1978

[15] Keh-Yih Su et.al., Speech Recognition using weighted HMM and subspace IEEE Transactions on Audio, Speech and Language.

[16] L.R.Bahl et.al, A method of Construction of acoustic Markov Model for words, IEEE Transaction on Audio ,speech and Language Processing ,Vol.1,1993

[17] Shigeru Katagiri et.al., A New hybrid algorithm for speech recognition based on HMM segmentation and learning Vector quantization , IEEE Transactions on Audio Speech and Language processing Vol.1,No.4

[18] G. 2003 Lalit R .Bahl et.al.,Estimating Hidden Markov Model Parameters so as to maximize speech recognition Accuracy,IEEE Transaction on Audio, Speech and Language Processing Vol.1 No.1 , Jan.1993.

[19] Mari ostendorf et.al. from HMM to segment Models: a Unified View stochastic Modeling for speech Recognition ,IEEE Transaction on audio, speech and Language Processing Vol.4,No.5,September 1996.

[20] John butzberger ,Spontaneous speech effects In Large Vocabulary Speech Recognition application,SRI International Speech Research and Technology Program Menlo Park,CA 94025

[21] Dannis Norris, "Merging Information in Speech Recognition" feedback is never Necessary workshop.1995

[22] Yifan gong, stochastic trajectory Modeling and Sentence searching for continuous Speech Recognition,IEEE Transaction on Speech and Audio Processing,1997.

[23] Alex weibel and Kai-Fu Lee, reading in Speech recognition ,Morgan Kaufman Publisher,Inc.San Mateo,California,1990.

[24] John Butzberger, Spontanious Speech Effect in Large Vocublary speech

recognition application, SRI International Speech Research and Technology program Menlo Park, CA94025.

[25] E. Singer, P.A. Torres-Carrasquillo, T.P. Gleason, W.M. Campbell, and D.A. Reynolds ,"Accoustic ,phonetic and discriminative approach to automic Language Idantification".

[26] Viet Bac Le, Laurent Besacier, and Tanja Schultz, Acoustic- phonetic unit similarities for context dependant acoustic model portability Carnegie Mellon University, Pittsburgh, PA, USA

[27] C.S.Myers and L.R.Rabiner, A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition , IEEE Trans. Acoustics, Speech Signal Proc.,ASSP-29:284-297,April 1981.

[28] D.R.reddy,An Approach to Computer speech Recognition by direct analysis of the speech wave,Tech.Report No.C549,Computer Science Department ,Stanford University,sept.1996