JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Study of speech recognition based on Different Deep Learning techniques: A Review

Kousik Mallick
Department of Artificial Intelligence
Reva University, Bangalore

Abhijit Mallick
Wipro CTO innovation labs

Abstract*—Speech is the most effective technique for people to transmit their thoughts and feelings across a multitude of languages. Every language has a particular set of speech characteristics. The speed and accent differ from person to person even while speaking the same language. This might make it hard for certain people to understand the point being made. Long speeches may be difficult to follow if the speaker has poor pronunciation, speaks too quickly, or uses other techniques to lose the audience's attention. Speech recognition is an interdisciplinary topic of computational linguistics that enables the creation of technology that recognises and transcribes sound into text. The most relevant information is collected from a text source and appropriately summarized via text summarizing

Keywords*— Speech recognition, noisy speech recognition, deep learning.

i. INTRODUCTION

Humans have an innate ability to communicate via speech. All the necessary knowledge is acquired automatically throughout early infancy, and we depend heavily on verbal exchanges all through our lives. We take language for granted and fail to appreciate its complexity because of how easily it comes to us. Human vocal tract and articulators are nonlinear biological organs whose function is influenced by a wide variety of circumstances, including but not limited to gender, upbringing, and emotional state. As a result, there may be significant variations in vocalisations with respect to accent, pronunciation, articulation, roughness, nasality, pitch, loudness, and speed; furthermore, our already erratic speech patterns may be further affected during transmission by things like background noise, echoes, and electrical properties. Because of all these different factors, speech recognition is significantly more difficult than speech creation.

ii. OBJECTIVES OF SPEECH RECOGNITION

Many handicapped persons benefit greatly from speech recognition since it enables them to use a wide range of gadgets and appliances without having to use their hands. Some of the early uses of voice recognition technology were in medical dictation software and automated telephone systems [2]. Each part of a speech recognizer—from the audio input to the feature extraction to the feature vectors to the decoder to the final word output—is essential. The decoder makes use of language models, pronunciation dictionaries, and acoustic models. Benefits:

- Many sectors, including the healthcare sector, might benefit from its use to boost productivity.
- It can record voice at a far higher rate than typing.
- The text-to-speech feature works in real time.
- In terms of spelling, the programme is on par with any other word processor. Beneficial for persons with hearing or visual impairments.

iii. AUTOMATIC SPEECH RECOGNITION

Automatic speech recognition is the simple understanding by a machine of human speech and the subsequent realisation that the human voice directly commands the computer, which then acts on those commands based on the recognised and processed voice, realising the intelligent interaction between human and computer.Most of the traditional automatic speech recognition models use Hidden Markov Gaussian Mixture Model (HMM-GMM)[1], which is based on the theory of likelihood and probability. Figure 1 depicts the framework of the more complicated and multi-participant classical speech recognition model. In the front-end speech preprocessing step, the standard automatic speech recognition model uses the connection between the speech signal and the digital model to make predictions about the signal; to do this, it combines samples of the signal and use

linear prediction analysis. However, it is challenging to adapt the preprocessing models used in conventional speech recognition to these settings owing to the complexity of speech information extraction caused by the varying pronunciations of persons of various languages, genders, and ages.

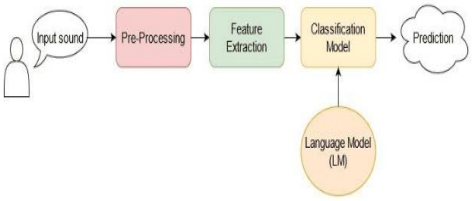

Figure 1: automatic speech recognition

Types of automatic speech recognition

The limits placed on the input speech allow for the classification of voice recognition systems into several classes.[2].

- **Number of speakers:** The term "speaker independent" refers to a system's ability to recognise the speech of any speaker, regardless of the speaker's individual characteristics. Training a speaker-dependent system requires a vast quantity of speech data from a user.
- **Nature of the utterance:** In order for an Isolated Word Recognition system to work, the user must speak in clearly separated chunks of text. Without pausing in between sentences, a Connected Word Recognition system can identify specific words from a limited vocabulary. Conversely, Continuous Speech Recognition apparatuses identify constantly uttered phrases.
- **Vocabulary size:** An ASR system with a small vocabulary can only recognise a handful of words, maybe only ten digits. Systems with a medium vocabulary may understand a few hundred words at most. Both Large and Very Large ASR systems need training on tens of thousands of words.
- **Spectral bandwidth**: The frequency components outside of the 300–3400 Hz passband are attenuated by the narrow bandwidth of the telephone/mobile channel. Narrowband speech describes this kind of communication. Wideband speech, on the other hand, is regular speech that bypasses that channel and includes a larger spectrum limited only by the sampling frequency.

iv. FUNDAMENTAL OF SPEECH RECOGNITON

The process of speech recognition involves analysing and organising acoustic data into a hierarchy of subword units, words, phrases, and sentences. Additional temporal restrictions, such as recognised word pronunciations or legal word sequences, may be provided at higher levels to account for faults or uncertainty at lower levels. Taking advantage of this hierarchy of restrictions requires making discrete choices at the top level while mixing probabilistic choices at the lower levels. Figure 2.1 depicts a typical speech recognition system's layout. The components consist of:

1) **Raw speech:** In the case of a microphone, 16 kilohertz (KHz), and in the case of a telephone, 8 kHz, the frequency at which speech is collected is rather high. This results in a time series of amplitude readings.
2) **Signal analysis**: It is recommended to first convert and compress raw voice before further processing. There are a plethora of signal analysis methods that may be used to extract meaningful characteristics and compress the data by a factor of ten without sacrificing any essential details.

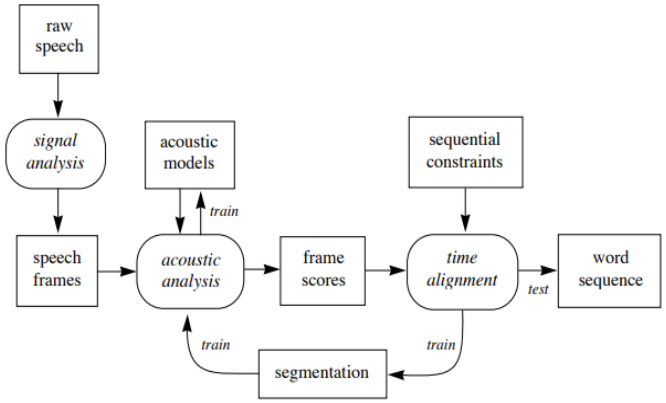

Figure 2: structure of speech recognition system

v. KEY FEATURES OF EFFECTIVE SPEECH RECOGNITION

There are a variety of voice recognition apps and gadgets out there, but the most cutting-edge options use AI and machine learning. To decipher human speech, they combine linguistic concepts like as grammar and syntax with information on the structure and composition of audio and voice signals. They should be able to adapt their behaviour based on what they've learned from previous experiences. The most effective solutions also provide businesses the option to tweak the software to better suit their needs, from language and voice subtleties to logo recognition. For instance:

- **Language weighting:** Accuracy may be enhanced by giving additional importance to frequently used words beyond those already included in the basic vocabulary.
- **Speaker labeling:** Generate a transcript of a group discussion that includes references to or tags for each participant's comments.
- **Acoustics training:** Look after the noise levels at work. To make the system work better in noisy settings (like a call centre) and accommodate different speaking styles, we must first train it to do so (like voice pitch, volume and pace). Filtering for obscenities involves analysing input speech for certain words or phrases in order to clean it up before it's broadcast.

vi. BASIC MODEL OF SPEECH RECOGNITION

People's motivation to study voice and language was mostly driven by their ambition to create mechanical models of human verbal communication. Due to the inherent advantages of spoken language, speech processing has emerged as a promising branch of signal processing. The development of speech recognition technology has allowed computers to obey vocal directions and translate across languages. The primary focus of the voice recognition field is the creation of methods and systems through which a computer can receive input from human speech. Humans rely mostly on verbal exchanges for communication. A lot of people have been interested in researching automatic speech recognition by machines for the past sixty years for a variety of different reasons, including a technological curiosity about the mechanisms for mechanical realisation of human speech capabilities and a desire to automate simple tasks that require interactions between humans and machines.

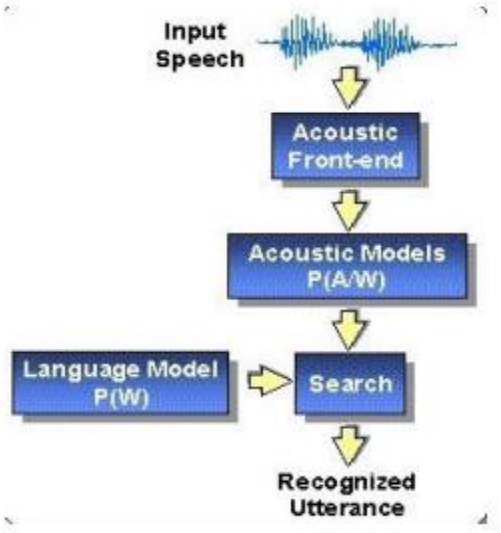

Figure 3: model of speech recognition

vii. DEEP LEARNING

Deep learning is a subfield of machine learning that involves multilayered neural networks. These neural networks "learn" from extensive datasets in an effort to mimic human brain activity, albeit they are still far from brain supremacy. Although approximations might still be produced by a neural network with just a single layer, the addition of hidden layers can aid in optimization and refining, resulting in more accurate predictions. Many AI apps and services rely on deep learning to boost automation by handling analytical and physical activities that previously required human participation. Both established and up-and-coming technologies rely on deep learning techniques, including digital assistants, voice-enabled TV remotes, and credit card fraud detection (such as self-driving cars). Speech recognition utilises a variety of deep learning approaches.

1) Convolutional neural network

The primary applications of the neural network subtype known as convolutional neural networks are in the areas of image classification, picture clustering, and object identification. By using DNNs, hierarchical visual representations may be built without human supervision. Deep convolutional neural networks are favoured over other types of neural networks because they provide the highest levels of accuracy.

2) Recurrent Neural Network (RNN)

Using sequential or time-series data, the result of a previous stage is fed as input to the current stage in a RNN, another kind of popular neural network. [3]. Recurrent networks, like feedforward and CNN, learn from training input, but they are distinguished by their "memory," which influences both input and output in the present by drawing on information from the past. The output of an RNN depends on the components that came before them in the sequence, which is not the case for a standard DNN, which believes that inputs and outputs are unrelated.

3) Restricted Boltzmann Machine (RBM)

The RBM may be thought of as a Boltzmann Machine. In this setup, the input layer neurons and the hidden layer neurons are connected in a symmetrical fashion. However, inside that specific layer, no such link exists. Boltzmann machines, unlike RBM, have hidden-layer internal connections.[4]. These BM limitations aid the model's learning process.

4) Recurrent Neural Network

One additional subset of feed-forward networks are recurrent neural networks. In this case, all of the neurons in the hidden layers get their inputs at different times. To a large extent, the current iteration's data is stored in the past for use by the Recurrent neural network. If you want to predict the next word in a sentence, you need to know what came before it.[3]. It does more than just process inputs; it also distributes length and weights laterally through time. It prevents the model size from becoming proportionally larger when more data is added. This recurrent neural network has the single drawback of being computationally sluggish and not considering any future input for the present state. It has trouble recalling previously learned material.

5) Deep Belief Network (DBN)

A DBN [5]refers to a kind of multi-layer generative graphical model built by feeding the output of one unsupervised network (such an AE or RBM) into the input of another.[6]. Input data with a deep structure may be represented hierarchically when using DBN.

6) Autoencoders

In the realm of machine learning, the autoencoder neural network is another kind of unsupervised method. In this case, the fraction of input cells that remain concealed is rather tiny. However, the output number is always the same as the input number of cells. In order to get AEs to detect common patterns and generalise the data, a network of autoencoders is trained to produce output that looks like the fed input. [7]. Autoencoders are often used to reduce the size of the input representation. When applied to compressed data, it aids in reassembling the original. In order for this algorithm to work, all it needs is for the output to be the same as the input.

These many applications of deep learning may be roughly categorised into four types: (a) computer vision; (b) voice recognition; (c) natural language processing (NLP); and (d) recommendation engines.

7) Computer vision

The term "computer vision" refers to a computer's capacity to analyse visual data such as photographs and movies. With the use of deep learning methods, computers can understand pictures just as well as humans. The following are some examples of the many uses for computer vision:

- The ability to have media collections automatically scanned for harmful or objectionable content.
- Face recognition based on characteristics such as hair, glasses, and eye opening
- Visual recognition of trademarks, garments, and protective equipment

8) Speech recognition

Human speech may be analysed by deep learning models regardless of language, accent, or pitch. The following are examples of the uses of voice recognition in virtual assistants like Amazon Alexa and automated transcribing software:

- Assist call center agents and automatically classify calls.
- Real-time recording of clinical discussions.

- Subtitrate films and audio recordings of meetings with precision to increase their accessibility.

9) Natural language processing

Deep learning methods allow computers to extract information and meaning from large amounts of text data and texts. There are several applications for being able to process human-created natural language text.

- Automated virtual agents and chatbots
- Documents and news stories may be automatically summarised.
- Long-form content analysis for business purposes. This includes but is not limited to emails and forms.
- Sentiment analysis via the indexing of key terms indicating favourable and negative feedback from social media platforms

viii. LITERATURE REVIEW

Speech is the primary means through which humans interact with one another. Speaking is typically considered the main means of communication, despite the fact that there are many other ways to convey our thoughts and feelings. Recorded voice may be transcribed with the help of the Google API. Retrieving text without a period makes it difficult to divide it into sentences for usage with the Google API. A period is placed after each sentence in the proposed structure to denote their individuality.

this researcher [2] selected the required audio from a huge file and listened to it. This study used deep learning to categorise vocalisations. The model was trained using data from Google's corpus. There was a 66.22 percent success rate.

In this study [8], A comprehensive knowledge of SER has formed via the use of deep learning techniques, such as the preprocessing of audio signals, feature extraction and selection methods, and the last step of proving the accuracy of a good classifier. Emotional data from Ravdess, Crema-D, Tess, and Savee were used to fine-tune the training of the one-dimensional Convolutional Neural Network. On the combined Ravdess, Crema-D, Tess, and Savee datasets, the feature combination of ZCR+energy+entropy of energy+RMS+MFCC achieved 92.62 percent accuracy.

In this study [9] Examine the speech processing flow of a voice recognition-based system for teaching and learning a foreign language. Codebook production and template training then make use of HMM-based voice recognition technology for feature parameter extraction, while the enhanced Viterbi model is employed to cut down on Gauss calculation. Last but not least, the expert database is utilised to fix phonemes, as shown via real-world spoken English instruction. The suggested technique increases recognition accuracy by 15% using real-world data from massive-scale spoken English assessments, allowing for more fast, accurate, and objective evaluation and feedback advice for students.

In this study [10] an innovative strategy emerges, harmonizing the robust CNN model with MFCC feature extractions, to unveil the latent emotional intricacies within speech data. Information from many databases was merged to improve the voice samples. The mel-spectrograms (static, delta, delta, and delta) retrieved from the audio signal were used as input to a CNN model. CNN retrieved segment-level features, and utterance-level characteristics were gathered, and then layered. The completed Speech Emotion Recognition (SER) system made use of the CNN model. Popular SER (Speech Emotion Recognition) is used in the suggested technique, and it has done better than previous methods.

In this work, [12] With deep reinforcement learning, you may offer a new method for automatically determining when to cease rechecking inferior ASR transcriptions. When compared to the baseline system, which just considers the best ASR result, our approach is 3.1% more effective. Our strategy may increase DST accuracy by 15.4%, which is five times the total improvement rate, by selecting the conversation rounds with the top-10 biggest word error rate (WER) (3.1 percent). As we hypothesised, this enhancement would materialise thanks to our suggested method's ability to pick out useful ASR findings at any rank.

The aim of this study [13] is to find out which approach to deep learning will work best for a pilot project. The CNN model and the LSTM model were used to train deep-learning neural networks on a very modest voice corpus. The models were compared and assessed using a cross-validation method. For a limited amount of data, CNN fared better than LSTM. Given the research's focus on voice recognition and the dataset and architecture size constraints, this seems to indicate that CNN is the best option.

ix. CONCLUSION AND FUTURE WORK

Voice processing tasks have benefited greatly from the fast development of deep learning methods, with major leaps forward in areas like speech recognition, speaker identification, and speech synthesis. This study offers a thorough analysis of recent advances in the application of deep learning methods to problems involving voice processing. We start with a brief history of voice processing, covering topics like representation learning and HMM-based modelling, then go on to a comprehensive overview of core deep learning methods and their implementations in speech processing. We also cover the most important problems in voice processing, emphasise the datasets that are employed, and provide the most recent and significant research studies that have applied deep learning.

x. REFERENCES

[1] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech Recognition Using Deep Neural Networks: A Systematic Review," *IEEE Access*, 2019, doi: 10.1109/ACCESS.2019.2896880.

[2] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," *Multimed. Tools Appl.*, 2021, doi: 10.1007/s11042-020-10073-7.

[3] C. J. Mandic D, "Recurrent neural networks for prediction: learning algorithms, architectures and stabilit," *Hoboken: Wiley*, 2001. .

[4] B. M. Marlin, K. Swersky, B. Chen, and N. De Freitas, "Inductive principles for restricted Boltzmann machine learning," 2010.

[5] H. GE., "Deep belief networks," *Scholarpedia*, 2009. .

[6] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, 2006, doi: 10.1162/neco.2006.18.7.1527.

[7] A. Ben Miled, T. Yeferny, and A. Ben Rabeh, "MRI Images Analysis Method for Early Stage Alzheimer's Disease Detection," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 20, no. 9, pp. 214–220, 2020, doi: 10.22937/IJCSNS.2020.20.09.26.

[8] S. Ullah, Q. A. Sahib, Faizullah, S. Ullahh, I. U. Haq, and I. Ullah, "Speech Emotion Recognition Using Deep Neural Networks," in *2022 International Conference on IT and Industrial Technologies (ICIT)*, 2022, pp. 1–6, doi: 10.1109/ICIT56493.2022.9989197.

[9] X. Chu, "Speech Recognition Method Based on Deep Learning and Its Application," in *2021 International Conference of Social Computing and Digital Economy (ICSCDE)*, 2021, pp. 299–302, doi: 10.1109/ICSCDE54196.2021.00075.

[10] A. Arul Edwin Raj, K. K. B, S. S, and R. A, "Speech Emotion Recognition using Deep Learning," in *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, 2023, pp. 505–509, doi: 10.1109/ICIDCA56705.2023.10100056.

[11] P. Singh, M. Sahidullah, and G. Saha, "Modulation spectral features for speech emotion recognition using deep neural networks," *Speech Commun.*, 2023, doi: 10.1016/j.specom.2022.11.005.

[12] R. T.-H. Tsai, C.-H. Chen, C.-K. Wu, Y.-C. Hsiao, and H. Lee, "Using Deep-Q Network to Select Candidates from N-best Speech Recognition Hypotheses for Enhancing Dialogue State Tracking," in

ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 7375–7379, doi: 10.1109/ICASSP.2019.8683749.

[13] J. Kaewprateep and S. Prom-On, "Evaluation of small-scale deep learning architectures in Thai speech recognition," 2018, doi: 10.1109/ECTI-NCON.2018.8378282.