IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 5, May 2022

Deep Learning Based Speech Recognition: A Review Paper

Rohaila Naaz, Assistant Professor,
College of Computing Sciences and Information Technology, Teerthanker Mahaveer University,
Moradabad, Uttar Pradesh, India
Email Id- rohailanaaz2@gmail.com

ABSTRACT: ***The act of converting audible words into written ones is called speech recognition. To recognize speech, sound waves must be captured and digitally transformed into basic digital representations. Examining the contexts and phonemes of words to ensure the correct spelling for words with comparable sounds is key when creating words from linguistic components or phonemes. The purpose of the paper is to review pattern matching and deep learning-based speech recognition methods explored in recent years. Intelligent virtual assistants or intelligent personal assistants are terms used to describe software agents that carry out tasks or offer services on behalf of users. Private supported commands or queries are very essential in modern speech recognition systems. When the phrase is exclusively accessed through online chat or asking generic questions of virtual assistants, then the online chat programs can occasionally simply be used for fun.***

KEYWORDS: ***Deep Learning, Machine Learning, User Speech, Neural Networks, Chat, Software.***

1. INTRODUCTION

Speech recognition is the ability of a computer or program to understand words and phrases in spoken language and transform them into a machine-readable format. There are currently several speech recognition applications available, such as voice calling, basic data entry, and speech-to-text. Systems for automatic voice recognition use many different components that are drawn from several statistical pattern recognition disciplines, such as combinatorial, signal processing, communication theory language, and mathematics. Speech synthesis is a replacement for traditional communication techniques, such as typing text into a computer using a keyboard. An effective system can replace or reduce the reliance on traditional keyboard input, enabling autonomous voice recognition (ASR) the earliest systems emerged in the 1950s. As computers, mobile devices, and other electronic gadgets penetrated our daily lives, they used both simple and complex systems to reduce monotonous work and the waste of human resources [1], [2]

Virtual personal assistants are almost universally regarded as the absolute minimum. To conveniently carry out the necessary tasks on all electronic devices. VPA is capable of more than just acting as a bot; it can also make things better. The user in many different ways. One is speech synthesis. Modern concepts like neural networks and machine learning are used in the complex process of voice recognition in speech-to-text. The auditory input is processed by a neural network utilizing vectors. Is designed to fit each letter and phrase. It is what it is, the data set. The system evaluates a user's speech against this vector [3]–[6]. To produce deep learning, a variety of machine learning algorithms are fed inputs in the form of many-layered models. These models frequently include neural networks with different non-linear operating levels. The gadget Learning algorithms make an effort to learn from these deep neural networks by omitting particular data and attributes. Deep architecture inputs could not be searched before 2006. The simple and predictable task, yet with the help of deep learning techniques, this issue was resolved by scanning the parameter space of the condensed architectural depths. In speech-

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 5, May 2022

to-text, deep learning models may also be used. As an unsupervised layer-by-layer pre-training that is greedy.

Fewer parameters are required to express a, according to studies. Compared to the several parameters needed to represent the same function with a shallower architecture, a given non-linear function in a deep architecture. Deep learning methods are mostly used to enhance computer capabilities so that they can understand what humans can do, including speech recognition. Being the main mode of communication amongst people. Therefore, it makes sense that speech was one of the first applications of deep learning, and several studies have been conducted and published to this day on the subject. Used deep learning for tasks involving speech-to-text in publications applications that place a strong emphasis on speech recognition. Contrary to HMMs, neural networks significantly improve discriminative training [7]–[11]. Figure 1 illustrates the deep learning model for speech recognition.

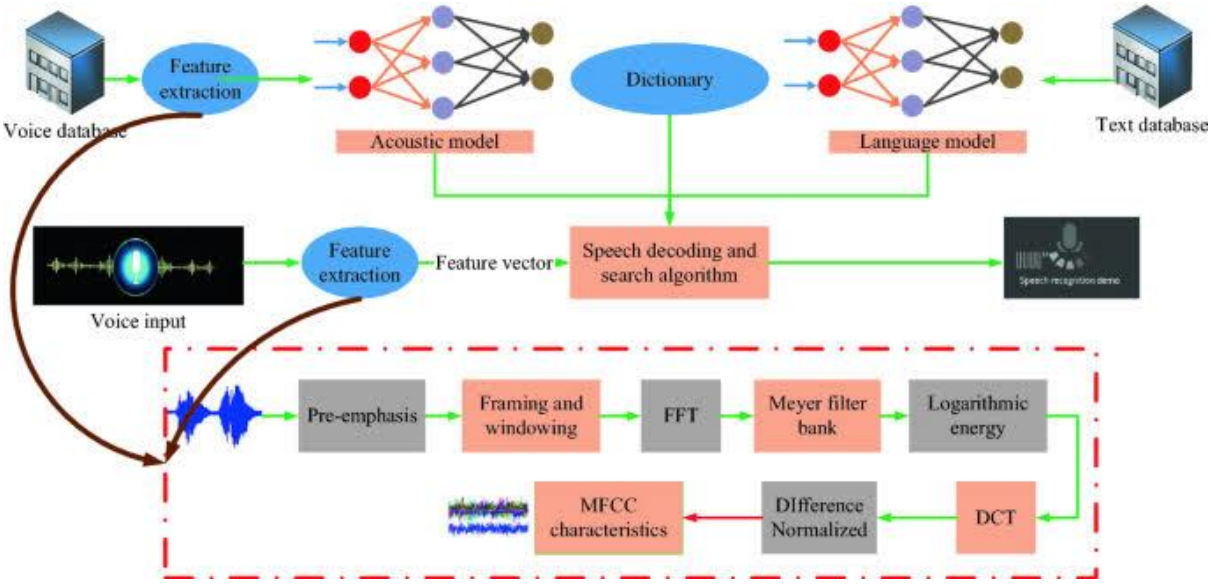

Figure 1: Illustrates the deep learning model for speech recognition [Google].

The current ASR systems make use of intricate statistical models. The success of hidden Markov models has been enormous. These models generate a string of symbols or numbers using statistics. There are GMM-HMMs used. A voice signal may be viewed as a transitory or piecemeal fixed signal, depending on the context. The proposed system will use "learning" algorithms that try to directly comprehend the features. Recently, pattern recognition problems have seen great success using neural network-based approaches. Mostly as a result of increased computational power. In contrast Neural networks and GMM-HMMs make no assumptions and offer a wide range of statistical properties that make them superior voice synthesis recognition models. An amalgamation of the depth of the internet and the phone's accessibility and mobility are increasingly forming a large part of society [12].

Converting spoken words into text is a task that is done in computer science and electrical engineering. The terms "computer speech recognition," "automatic speech recognition," and "speech to text" are also used to refer to it (STT). Other SR systems use "training," in which a single speaker reads text fragments to the SR system. Some SR systems use "speaker-independent speech recognition," while others use "training." These systems analyze the subject's speech and use it to enhance the subject's speech recognition, resulting in a more

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 5, May 2022

accurate transcription. Training systems are not necessary for "speaker-independent" systems. Systems that rely on training are known as "speaker-dependent" systems. Human behavior, particularly the capacity for communication and natural behavior, engineers and scientists are

2. DISCUSSION

The fundamental idea of speech is that human sounds are filtered by the morphology of the vocal tract, including the tongue, teeth, and other structures. The sound that is produced is determined by this form. If we can understand this, it should provide us with a clear picture of the phoneme that is being formed. The vocal tract's shape exhibits signs of the vocal short temporal power spectrum, and it is the responsibility of MFCCs to accurately represent this envelope. The perceptron is the most fundamental supervised learning feedforward network. A perceptron is made up of binary threshold units. Arranged in layers. Theoretically, multi-layer perceptrons, or MLPs, are capable of learning any function, but they are more challenging to train. The Delta Rule does not apply. Since there Extreme parallelism is made possible by the concurrent operation of these complete units. All system computing is handled by these components; no other processor is involved in their management or execution. Each unit simply behaves at the moment as it does. Computes a scalar function based on its local inputs and transmits the result, also referred to as the activation value, to its neighboring units. The units of a network are often divided into input units, which receive information from the environment (such as raw sensory data), cryptic units, which have the potential to change the data internally, and output units, which either reflect judgments or command messages. Three samples covering the range of 0 to 25 milliseconds were taken. The phoneme activations are sent to the word and syntax recognition part of the recognition system, which uses them.

Up until recently, the majority of signal processing techniques made use of shallow organized structures. The number of non-linear feature modification layers in these architectures is often no more than one or two. Examples of these shallow architectures are support vector machines and Gaussian combination models (SVMs). The idea of deep learning was initially inspired by artificial network research. A good illustration of a model with a deep architecture is a deep neural network, often known as a feed-forward neural network. Backpropagation was one of the most used algorithms for determining these networks' parameters (BP). However, when BP was employed exclusively, learning networks had trouble. They have quite a few layers that are buried. Recurrent occurrence of local optimum in non-convex objective operations of deep networks. Computing a sparse set of feature vectors that gives a concise representation of the input audio is the primary goal of the feature extraction stage in voice recognition. The feature extraction is normally finished in three stages. Figure 2 illustrates the process of deep learning in speech recognition.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 5, May 2022

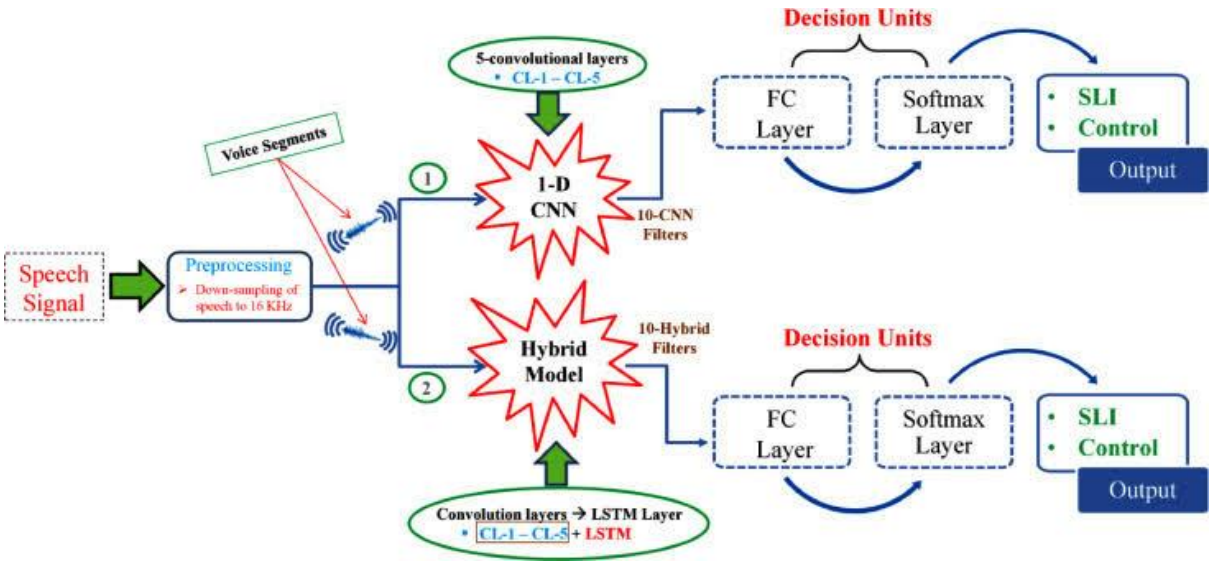

Figure 2: Illustrates the Process of Deep Learning in Speech Recognition [Google].

The first step is speech analysis, also referred to as the acoustic front end. For brief speech intervals, it does some sort of spectrum temporal analysis on the signal and generates raw features that define the power spectrum's envelope. The second stage involves assembling an extended feature vector made up of both static and dynamic features. Finally, more dependable and manageable feature vectors are created from these extended feature vectors and provided to the recognizer. You may develop a classifier. One could ponder whether sufficient ASR will ever be fully attained. Most artificial intelligence (AI) tasks are viewed as potentially realizable; unquestionably, considerable breakthroughs in robotics and chess-playing machines support this claim. In contrast to the task of driving a car without human assistance, the latter calls for intelligence in the interpretation of the field of vision of a mounted camera on a car. However, the algorithms needed for cars share signal characteristics that are very unlike ASR. Processing feels overwhelming, as do both challenges (i.e., substituting a human driver with a similarly capable algorithm may appear to be as unrealistic as having a fully recognizing ASR device). However, it seems that ASR is a workable solution that is much closer.

Shown that cutting-edge outcomes in phoneme recognition on the TIMIT database can be achieved by combining deep, bidirectional Long Short-Term Memory RNNs with end-to-end training and weight noise. Extending the technology to support large vocabulary speech recognition is a logical next step. Combining deep LSTM with frequency-domain convolutional neural networks is another intriguing direction. Since RNNs are flexible, regularisation is essential for excellent performance because overfitting is a risk with these models. Early halting and weight noise (the insertion of Gaussian noise to the network weights) were the two regularizes utilized.

3. CONCLUSION

In this study, familiar deep learning techniques like deep neural networks (DNN) and deep belief networks (DBN) have been examined and understood for speech recognition. Additionally, a DBN has been set up for automatic voice recognition. The TIMIT acoustic-phonetic continuous speech corpus dataset rate was used to rate the word mistake rates of the three voice recognition systems, GMM-HMM, DNN-HMM, and DBN. The findings of this study show that the DBN-based voice recognition system works better than the other two. One

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 5, May 2022

could ponder whether sufficient ASR will ever be fully attained. Most artificial intelligence (AI) tasks are viewed as potentially realizable, unquestionably, considerable breakthroughs in robotics and chess-playing machines support this claim. Recall that surrounding portions are sensitive, which could be because CTC tries to infer linguistic connections from the auditory input while not having an explicit language model.

REFERENCES:

[1] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," *IEEE Access*, 2019, doi: 10.1109/ACCESS.2019.2936124.

[2] D. Wang, X. Wang, and S. Lv, "An overview of end-to-end automatic speech recognition," *Symmetry*. 2019. doi: 10.3390/sym11081018.

[3] L. Liu *et al.*, "Deep Learning for Generic Object Detection: A Survey," *Int. J. Comput. Vis.*, 2020, doi: 10.1007/s11263-019-01247-4.

[4] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electron. Mark.*, 2021, doi: 10.1007/s12525-021-00475-2.

[5] J. M. Ede, "Deep learning in electron microscopy," *Machine Learning: Science and Technology*. 2021. doi: 10.1088/2632-2153/abd614.

[6] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, 2019, doi: 10.1186/s40537-019-0192-5.

[7] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," *ACM Computing Surveys*. 2019. doi: 10.1145/3285029.

[8] C. Cao *et al.*, "Deep Learning and Its Applications in Biomedicine," *Genomics, Proteomics and Bioinformatics*. 2018. doi: 10.1016/j.gpb.2017.07.003.

[9] A. Muniasamy and A. Alasiry, "Deep learning: The impact on future eLearning," *Int. J. Emerg. Technol. Learn.*, 2020, doi: 10.3991/IJET.V15I01.11435.

[10] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Computers and Electronics in Agriculture*. 2018. doi: 10.1016/j.compag.2018.02.016.

[11] S. A. Bello, S. Yu, C. Wang, J. M. Adam, and J. Li, "Review: Deep learning on 3D point clouds," *Remote Sensing*. 2020. doi: 10.3390/rs12111729.

[12] L. Deng, "Deep learning: From speech recognition to language and multimodal processing," *APSIPA Transactions on Signal and Information Processing*. 2016. doi: 10.1017/atsip.2015.22.