Speech Recognition Using Artificial Neural Network

C. Vivek[1,*], M. Indu[2], N. Nandhini[2]
[1]Department Of Artificial Intelligence and Data Science, Panimalar Engineering College, Chennai, India
[2]UG Scholar Department Of Artificial Intelligence and Data Science, Panimalar Engineering College, Chennai, India

Emails: vksundar7@gmail.com; indhumuthu1904@gmail.com; nandylefty0506@gmail.com

Abstract

Speech is a verbal communication used by humans through language. Likewise speech recognition is a process of converting speech to text. This paper provides a study of use of artificial neural networks(ANN) in speech recognition. Hidden Markov models (HMM) is a traditional statistical techniques for performing speech recognition. In speech detection software, Mel frequency cepstral coefficients (MFCCs) are frequently used. With different approaches evolving, we deal with the features used to recognize the speech pattern and implementation of speech recognition in the efficient types of artificial neural network (ANN).

Keywords: Artificial neural network, Signals; Artificial Neuron; Speech Recognition, Feature Extraction; Hidden layers; Artificial Intelligence, Hidden Markov Model(HMM),Mel Frequency Cepstrum Coefficients(MFCC)

1. Introduction

The process of translating spoken words into text is known as speech recognition[1]. It is a technique that enables computers to identify spoken language from humans and translate it into a machine-understandable format. Voice recognition systems use cutting-edge algorithms and artificial intelligence techniques to analyse and interpret the acoustic signals produced by human speech. These systems are designed to comprehend the sounds and patterns of speech even in noisy environments, whether the speaker has an accent, or if they have a speech disability. Dictation software, voice-activated virtual assistants, and voice-controlled interfaces for smart homes and transportation are few among the various uses for speech recognition technology[1].

2. Speech Recognition

Speech recognition is the process of converting spoken words into printed text or computer commands[2]. It is a type of artificial intelligence that enables computers and other objects to communicate with one another and grasp spoken language. Speech recognition technology blends machine learning algorithms with techniques from natural language processing to analyse and interpret spoken language. Voice recognition, language modelling, and acoustic processing are usually the three main stages.Acoustic processing comprises digitising the audio of spoken words after they have been recorded. These data are evaluated and transformed into a set of phonemes, which are the smallest units of sound in a language.

3. Speech Recognition Process

The speech recognition program works by scanning the speech, decodes it into bits that can be recognized, and then evaluates the content of each bit. The steps taken during the speech recognition process include

1.Signal Processing

Doi : https://doi.org/10.54216/JCHCI.050201
Received: December 19, 2022 Revised: March11, 2023 Accepted: May 01, 2023

2.Speech Enhancement
3.Feature Extraction
4.Acoustic Modelling
5.Phonetic unit Recognizing

3.1 Signal Processing

The first step in speech recognition is signal processing, which entails recording and altering the sound wave of human speech in order to transform it into a digital format that can be examined by a machine [4]. For precise speech recognition, this stage is essential. A microphone or equivalent device is used to record the sound wave before signal processing can begin. An analog-to-digital converter (ADC), which turns a continuous analogue signal into a series of discrete digital values, is then used to digitise the sound wave. After the sound wave is converted to digital form, it undergoes a number of processing steps to get rid of noise, distortion, and other artefacts that could obstruct accurate speech recognition.

3.2 Speech Enhancement

An important step in speech recognition is speech enhancement. Voice signals can be distorted in many different ways, such as channel distortion, reverberation, and noise, which might reduce the effectiveness of speech recognition programs. Speech Enhancement methods seek to eradicate these distortions, increasing the reliability and precision of systems that recognize speech [3]. Overall, speech augmentation is an important stage in the creation of speech recognition systems that aids in boosting the systems' robustness and precision under difficult conditions.

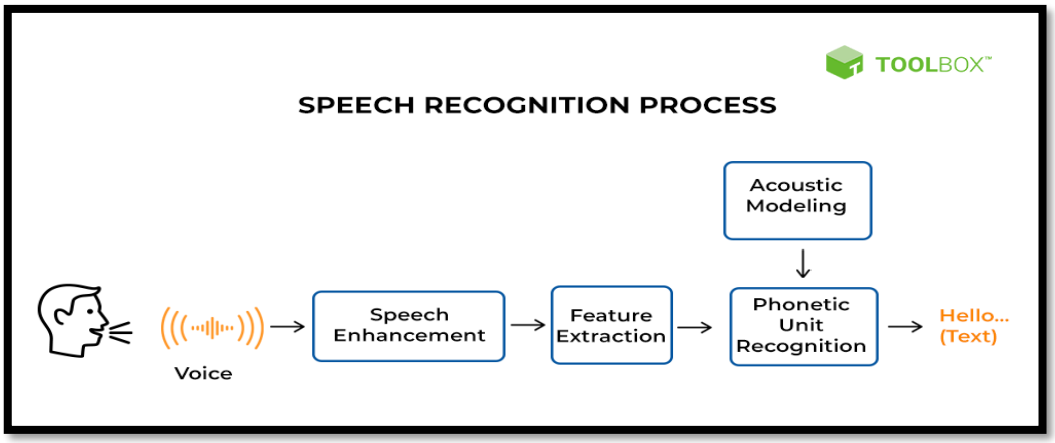

Figure 1: Speech Recognization Architecture

4. Feature Extraction

The analysis of the processed speech signal and the extraction of part of the analysis that can be utilized to distinguish and detect sounds of speech are essential steps in the speech recognition process. There are various steps involved in the feature extraction process, including:

4.1 Mel Frequency Cepstrum Coefficients (MFCC)

The use of mel frequency cepstral coefficients (MFCCs) is common in speech recognition software. They provide a compressed and decorrelated representation of the spectral envelope of spoken sounds. There are multiple processes in the MFCC feature extraction process[7]. The mel scale is a non-linear scale that simulates the frequency response of the human auditory system. It divides the frequency range into perceptually equal intervals. After filtering, the filterbank outputs are taken as a logarithm, and then a DCT (discrete cosine transform) is applied to decorrelate the resulting coefficients. Since they hold the majority of the speech data, the first few MFCCs typically 10–20 are kept for subsequent processing[8]. A crucial stage in the MFCC feature extraction process is a lot of speech recognition systems, and it aids in enhancing the reliability and accuracy of these systems in various settings and applications.

4.2 Linear Predictive Coding(LPC)

A common method for voice analysis and synthesis in speech recognition systems is linear predictive coding (LPC). LPC works by modelling the speech signal as a linear combination of previous signal samples, with

Doi : https://doi.org/10.54216/JCHCI.050201
Received: December 19, 2022 Revised: March11, 2023 Accepted: May 01, 2023

the coefficients of the linear combination being selected by an optimization process. The correlation between several speech signal samples is described by the autocorrelation function, which is employed[10]. The Levinson-Durbin algorithm and other methods are used to solve a set of linear equations to obtain the LPC coefficients. The speech signal can be synthesized with the help of the LPC coefficients once they have been obtained. LPC is frequently utilized in speech recognition systems. The effects of the vocal tract and linear distortion in the speech signal can also be removed with the help of LPC. In summary, LPC is a popular method for signal processing and feature extraction in speech recognition systems. It provides a compact representation of the speech signal[11].

5. Acoustic Modelling

Building statistical models of speech sounds or phonemes is known as acoustic modelling, and it is an essential part of automatic speech recognition (ASR) systems[5]. It can be difficult to capture speech variability in many settings and acoustic environments because speech signals have a high degree of fluctuation and noise. Training and decoding are the two processes that commonly make up acoustic modelling. The Hidden Markov Model (HMM) is the most popular type of statistical model used in ASR.

5.1 Hidden Markov Model(HMM)

A statistical model known as the Hidden Markov Model (HMM) is frequently employed in speech recognition systems to simulate the temporal fluctuation of voice signals.HMMs are used to represent the probability distribution of speech sounds[12]. HMMs are made up of a series of concealed states that result in a series of observable symbols. In order to determine the most likely sequence of hidden states that produced the observed symbols, the input voice signal is processed through the HMMs during the decoding phase. The temporal fluctuation of speech signals is modelled using HMMs, a statistical model that is frequently employed in speech recognition systems. With a sizable voice data, HMMs are taught to based on the input speech signal, estimate the model parameters and use them to determine the most likely sequence of hidden states during the decoding phase. In order to solve these issues and enhance the performance of speech recognition systems, researchers have created a number of extensions to HMMs[9]. HMMs have limits when it comes to modelling long-term dependencies and complicated interactions between speech sounds.

5.2 Phonetic Unit Recognition

The identification and decoding of the various phonemes that make up spoken words is a crucial part of automated speech recognition (ASR) systems. Training and decoding are typically the two phases of phonetic unit recognition[13]. During the training phase, the statistical models that are employed for phonetic unit recognition have their parameters estimated using a large corpus of speech data. Depending on the particular ASR system, these models might be Hidden Markov Models (HMMs), neural networks.The phonetic models are used to process the input speech signal during the decoding phase to generate a series of phoneme hypotheses. The decoding process involves searching through the space of possible phoneme to find the most likely sequence.In conclusion, phonetic unit recognition is an important part of ASR systems because it helps identify and decode the individual phonemes, that make up spoken words. [16-18]

6. Artificial Neural Network

A form of machine learning technique called an Artificial Neural Network (ANN) is motivated by the composition and operation of the human brain[4]. It is made up of a layer-organized network of interconnected nodes called neurons that are intended to process information similarly to how the human brain does.In an ANN, each neuron takes input from one or more other neurons, processes that input mathematically, and then generates an output signal that is transmitted to other neurons.Predictive modelling, speech recognition, natural language processing, image recognition, and other processes can all be done with ANNs.

Doi : https://doi.org/10.54216/JCHCI.050201
Received: December 19, 2022 Revised: March11, 2023 Accepted: May 01, 2023

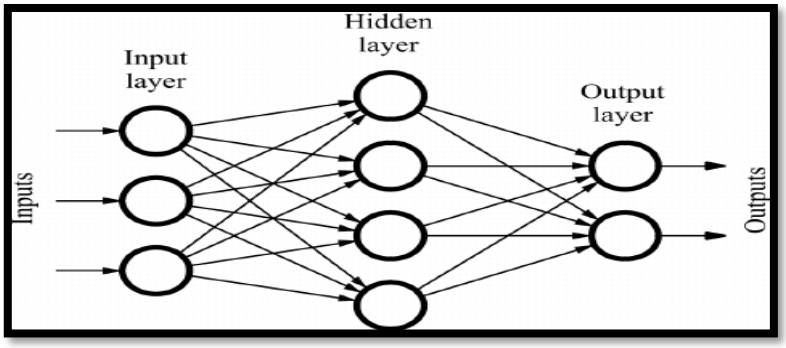

Figure 2:Artificial Neural Network**ARTIFICIAL NEURON:**

An artificial neuron is a basic building block of an artificial neural network (ANN). It is a mathematical function that takes in one or more inputs, performs a computation on those inputs, and produces a single output[14]. An artificial neuron typically consists of three parts: input weights, a summing function, and an activation function. The input weights are values that are multiplied by the input values to produce weighted sums. The summing function then aggregates the weighted sums to produce a single value. Finally, the activation function applies a non-linear transformation to the aggregated value to produce the output of the neuron.

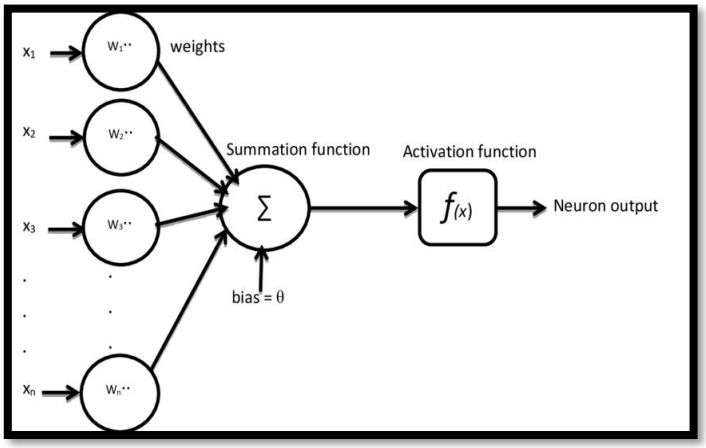

Figure 3**:** Artificial Neuron

7. Types Of Artificial Neural Netowrk

7.1 Feedforward Neural Network

A type of Artificial Neural Network (ANN) known as Feedforward Neural Networks (FNNs) is frequently utilized in speech recognition systems. FNNs are intended to take an information signal, like a discourse waveform, and produce a result signal FNNs take a sequence of audio features, like Mel-frequency cepstral coefficients (MFCCs), as input in a speech recognition system[1]. After that, these features are processed by the FNN through a series of hidden layers, with each layer transforming the input features in a nonlinear way. Lastly, a sequence of phoneme that match the input audio signal are produced by the FNNs output layer.A large amount of labeled speech data is used to train an FNN for speech recognition.Backpropagation is the most common method for training FNNs.

7.2.Recurrent Neural Network (RNN)

Due to their capacity to model temporal dependencies in sequential data, recurrent neural networks(RNNs) have found widespread application in speech recognition. RNNs maintain a hidden state that is updated at each time step based on the input and the last used hidden state. One famous sort of RNN for discourse acknowledgment is the Long Short-Term Memory (LSTM) organization[6]. A type of RNN called LSTMs has a memory cell that can store information over multiple time steps to better deal with long-term

Doi : https://doi.org/10.54216/JCHCI.050201
Received: December 19, 2022 Revised: March11, 2023 Accepted: May 01, 2023

dependencies.The input sequence is typically pre-processed in speech recognition to extract features like MFCCs,which are then fed into the LSTM network. Typically, minimizing a sequence loss function, such as the Connectionist Temporal Classification (CTC)loss, is part of training an RNN for speech recognition. This allows the network to learn to align the input sequence with the correct output sequence.RNNs, particularly LSTMs, have been extremely effective in speech recognition overall and are an essential component of numerous cutting-edge speech recognition systems[14].

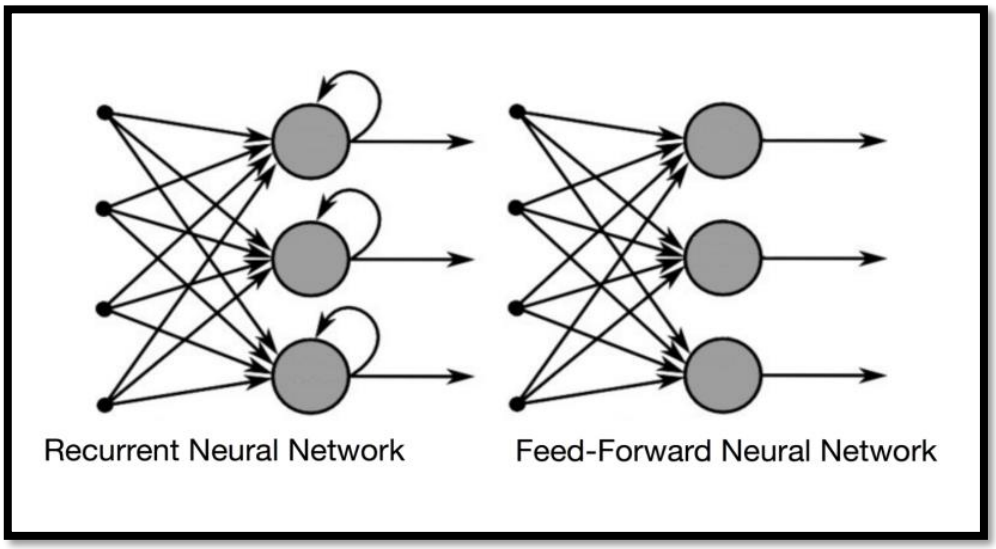

Figure 4: Recurrent Neural Network and Feed-Forward Neural Network

7.3.Modular Neural Network

An approach to the design of neural networks that consists of a number of smaller, simpler sub-networks or modules, each of which carries out a distinct function, are referred to as modular neural networks[12]. Modular architectures are another name for this type of design. By breaking down the complex task of speech recognition into smaller, more manageable subtasks, modular neural networks have been utilized in speech recognition to enhance the accuracy of the system. The hybrid neural network is one common modular architecture used in speech recognition. The neural network is used to estimate the likelihood of the acoustic features given the phoneme or word label, while the HMM is used to model the temporal dynamics of the speech signal.

The deep neural network (DNN) acoustic model is another modular architecture utilized in speech recognition. The DNN is typically trained with a large amount of labeled data to map the acoustic features to the phoneme or word label. The result of the DNN is then utilized as contribution to a Well, which models the worldly elements of the discourse signal. For speech recognition, end-to-end modular neural networks have recently been proposed. Convolutional neural networks (CNNs) for feature extraction and recurrent neural networks (RNNs) for sequence modeling are two examples of the sub-modules that these networks combine into a single end-to-end network. This strategy has the advantage of reducing the need for custom features and simplifying the training process. In general, speech recognition systems have improved in accuracy thanks to the success of modular neural networks in this field.

7.4 Kohonen Self Organizing Maps

Unsupervised neural networks like Kohonen Self-Organizing Maps (SOMs) can be used in speech recognition for things like speaker identification and phoneme clustering. SOMs function by mapping input data with high dimensions onto a grid with lower dimensions. Each neuron has an associated weight vector that is updated during training to represent a region of the input space. The neurons in the grid are arranged in a two- or three-dimensional lattice. The SOM modifies its weight vectors during training so that similar input patterns are mapped to nearby grid neurons. SOMs can be utilized in speech recognition to organize similar acoustic features, such as MFCCs or filterbank energies, into distinct SOM grid regions.This can be useful for automatic speech recognition and helping to identify phonetic features that are shared by multiple phonemes.

Doi : https://doi.org/10.54216/JCHCI.050201
Received: December 19, 2022 Revised: March11, 2023 Accepted: May 01, 2023

Advantages Of Artificial Neural Network

- *Data-based learning:* Because they can learn from data, ANNs are good for jobs where it is hard to specify clear rules or heuristics.
- *Generalization:* ANNs can make predictions based on data they haven't seen before because they can generalize to new inputs.
- *Non-linearity:* Because they are capable of capturing intricate, non-linear relationships between input and output variables, ANNs are well-suited for applications in which the underlying relationships are not well understood.
- *Tolerance for error:* Fault tolerance means that ANNs can continue working even if some of their parts break.
- *Processing in parallel:* For parallel processing, ANNs can be built into hardware or software, allowing them to process a lot of data quickly.
- *Adaptability:* ANNs are suitable for tasks where the input data may change over time because they can adapt to changing input patterns or environments.

Limitations Of Artificial Neural Network

- *Nature of a black box:* Because they are frequently treated as a black box with little understanding of the networks internal workings, ANNs can be challenging to interpret.
- *Overfitting:* Overfitting, in which the network learns the noise in the training data rather than the underlying patterns, can be a problem for ANNs. On new data, this may result in poor generalization performance.
- *Time for training:* Preparing an ANN can be computationally costly furthermore, tedious, particularly for enormous and complex organizations.
- *Choice of architecture:* In order to strike a balance between underfitting and overfitting, an ANN's architecture, including the number of layers and neurons, must be carefully selected. This can be a difficult task that may necessitate trial and error.
- *Requirements for data:* To learn meaningful patterns and adapt to new inputs, ANNs need a lot of training data.
- *Transferability restrictions:* If the problem domain and the input data are very different, ANNs trained for one task may not work well for another.

8. Conclusion & Future Scope

In speech recognition systems, both feedforward and recurrent neural networks (RNNs) have been used with success which depends on the requirements of the task and the characteristics of the input data.In practice, speech recognition systems have utilized both FFNNs and RNNs with success; which one to choose depends on the task specific requirements and the characteristics of the input data. A hybrid strategy that incorporates both kinds of networks may be utilized in some instances to capitalize on their advantages and overcome their disadvantages. The choice between FFNNs and RNNs for speech recognition ultimately depends on a number of factors, such as the system performance requirements, the available data, and the computational resources.

References

[1] Naziya Shaikh, Ratnadeep R. Deshmukh, "Speech Recognition System" 2016(DOI:10.9790/0661 1804020109)
[2] Suma Swamy,Kollengode Ramakrishnan "An Efficient Speech Recognition System" August 2013 (DOI:10.5121/cseij.2013.3403)
[3] https://www.guidogybels.eu/asrp4.html
[4] R. Subha Shini et.al., " Recurrent Neural Network based Text Summarization Techniques by Word Sequence Generation", IEEE International Conference on Inventive Computation Technologies (ICICT), 2021, DOI: 10.1109/I CICT50816.2021.9358764.
[5] Santosh K. Gaikwad, Bharti W. Gawali, Pravin Yennawar, "A Review on Speech Recognition Techniques", IJCA Vol. 10, No. 3, pp. 16-24, November 2010 (http://dx.doi.org/10.5120/1462-1976)
[6] Gasser Auda, Mohamed Kamel, "Modular Neural Network: A Survey", International Journal of Neural System, Vol. 9, No. 2, pp.129-151, April 1999.(http://dx.doi.org/10.1142/S0129065799000125)
[7] Shyam M. Guthikonda, "Kohonen Self-Optimizing Maps",Wittensberg University, December 2005.

Doi : https://doi.org/10.54216/JCHCI.050201
Received: December 19, 2022 Revised: March11, 2023 Accepted: May 01, 2023

[8] Premjeet Singh, Manoj Kumar Mukul,Rajkishore Prasad. "Chapter 14 Bone Conducted Speech Signal Enhancement Using LPC and MFCC", Springer Science and Business Media LLC, 2018
[9] Anjali I. P, Sherseena P. M "Speech Recognition"(DOI : 10.17577/IJERTCONV8IS04018)
[10] C. Gopala Krishnan,Y. Harold Robinson, Naveen Chilamkurti "Machine Learning Techniques for Speech Recognition using the Magnitude" (DOI : 10.33851/JMIS.2020.7.1.33)
[11] Vishnupriya Gupta "Speech Recognition Using Matrix Comparison" Sep-Oct. 2012(https://www.iosrjournals.org/iosr-jvlsi/papers/vol1-issue1/E0114345.pdf?id=1950)
[12] Sadeque, Zarif Al "Automatic Speech Recognition for Documenting Endangered First NationsLanguages"
[13] Chang, H.-C.. "Identification of lithofacies using Kohonen self-organizing maps",Computers and Geosciences, 2023
[14] Shaker K. Ali, Sabreen K. Saud. "Convert Arabic Letters Voice into Gesture, Journal of Physics: Conference Series, 2020
[15] Xiao JingHui, Liu BingQuan, Wang XiaoLong. "Chapter 72 Principles of Non-stationary Hidden Markov Model and Its Applications to Sequence Labeling Task", Springer Science and Business Media LLC, 2005.
[16] Mohammed K. Hassan , Ahmed K. Hassan , Ali I. Eldesouky, *MSJEP Classifier: "*Modified Strong Jumping Emerging Patterns" for Fast Efficient Mining and for handling attributes whose values are associated with taxonomies, Journal of Intelligent Systems and Internet of Things, Vol. 0 , No. 2 , (2019) : 37-53 (Doi : https://doi.org/10.54216/JISIoT.000201)
[17] Vijay K, Collaborating The Textual Reviews Of The Merchandise and Foretelling The Rating Supported Social Sentiment, Journal of Cognitive Human-Computer Interaction, Vol. 1 , No. 2 , (2021) : 63 - 72 (Doi : DOI: https://doi.org/10.54216/JCHCI.010203)
[18] P. Kavitha , R. Subha Shini , R. Priya, An Implementation Of Statistical Feature Algorithms For The Detection Of Brain Tumor, Journal of Cognitive Human-Computer Interaction, Vol. 1 , No. 2 , (2021) : 57 - 62 (Doi : DOI: https://doi.org/10.54216/JCHCI.010202)

Doi : https://doi.org/10.54216/JCHCI.050201
Received: December 19, 2022 Revised: March11, 2023 Accepted: May 01, 2023