Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2022.021755

check for updates

REVIEW

Challenges and Limitations in Speech Recognition Technology: A Critical Review of Speech Signal Processing Algorithms, Tools and Systems

Sneha Basak[1], Himanshi Agrawal[1], Shreya Jena[1], Shilpa Gite[2,*], Mrinal Bachute[2], Biswajeet Pradhan[3,4,5,*] and Mazen Assiri[4]

[1]Computer Science and Information Technology Department, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, 412115, India

[2]Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune, 412115, India

[3]Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), School of Civil and Environmental Engineering, Faculty of Engineering and Information Technology, University of Technology Sydney, New South Wales, 2007, Australia

[4]Center of Excellence for Climate Change Research, King Abdulaziz University, Jeddah, 21589, Saudi Arabia

[5]Earth Observation Centre, Institute of Climate Change, University Kebangsaan Malaysia, Bangi, Selangor, 43600, Malaysia

*Corresponding Authors: Shilpa Gite. Email: shilpa.gite@sitpune.edu.in; Biswajeet Pradhan. Email: biswajeet.pradhan@uts.edu.au

Received: 02 February 2022 Accepted: 30 June 2022

ABSTRACT

Speech recognition systems have become a unique human-computer interaction (HCI) family. Speech is one of the most naturally developed human abilities; speech signal processing opens up a transparent and hand-free computation experience. This paper aims to present a retrospective yet modern approach to the world of speech recognition systems. The development journey of ASR (Automatic Speech Recognition) has seen quite a few milestones and breakthrough technologies that have been highlighted in this paper. A step-by-step rundown of the fundamental stages in developing speech recognition systems has been presented, along with a brief discussion of various modern-day developments and applications in this domain. This review paper aims to summarize and provide a beginning point for those starting in the vast field of speech signal processing. Since speech recognition has a vast potential in various industries like telecommunication, emotion recognition, healthcare, etc., this review would be helpful to researchers who aim at exploring more applications that society can quickly adopt in future years of evolution.

KEYWORDS

Speech recognition; automatic speech recognition (ASR); mel-frequency cepstral coefficients (MFCC); hidden Markov model (HMM); artificial neural network (ANN)

1 Introduction

Researchers have been working on techniques to make computers capable of recording, interpreting, and understanding human speech since the 1960s. Speech and its interpretation appear to be simple, but in actuality, it is a complex motor ability due to which its mechanization interests many minds. With the rapid global development and investment in technology, system scope and power have

increased to a point where speech-driven systems can be taught in various industries like healthcare [1], telecommunication [2], e-commerce, education [3], and many more [2–4]. Speech recognition, in simple terms, is the ability of software or hardware to receive speech signals as input, analyze them, and accurately identify the words spoken correctly to execute a task based on them [5]. This technology is still evolving to add human-like functionalities like speaker identification, emotion recognition, gender identification, etc.

Data was gathered from research projects conducted between 1993–2021. The study includes 78 papers with Google Scholar as the search tool and the Web of Science (WoS) and Scopus databases as a reference for further details. Table 1 lists various surveys and review papers with detailed speech recognition analysis and its associated techniques, applications, and limitations. Most research papers seem to have focused on specific areas of speech signal processing, providing a summary of how various researchers have perceived and applied it through the decades. Although these papers provide a great deal of analysis of the journey of speech recognition systems, they often fail to provide a general approach that touches both the theoretical and practical aspects of creating an ASR [5–7]. On the other hand, this paper aims to fill the gap by touching on the basics of speech recognition systems and explaining the various approaches and tools needed for its implementation.

Table 1: Analysis of some speech recognition review papers

Paper	Content	Limitations
Feature extraction and classification techniques for speech recognition: A review [6].	This paper provides an overview of automatic speech recognition. It then examines the crucial topics such as types of speech classes, feature extraction techniques, speech classifiers, and performance evaluation.	This paper fails to provide a detailed analysis of the implementation tools and applications.
A study on automatic speech recognition [7].	This paper gives an overview of the background of Automatic Speech Recognition (ASR) as an essential domain of artificial intelligence. It puts forward the various types of classification of speech and the general architecture of speech recognition systems. It also summarizes crucial research relevant to speech processing and refers to specific enhancements in future works.	This paper talks about the two extremes of ASR systems, the basics of audio and the recent advancements. However, it lacks in mentioning the techniques and methodologies that have been used to create high accuracy systems in this domain.
A review on speech recognition [5].	This paper reviews the few primary techniques developed in each stage of creating a speech recognition system. It provides a comparative study of different techniques in the feature extraction stage and a detailed explanation of the traditional modeling techniques.	This paper fails to shed light on the development of speech recognition systems using the famous deep learning approach and gives a more theoretical overview of progress in speech analysis.

(Continued)

Table 1 (continued)

Paper	Content	Limitations
Speech recognition using deep neural networks: A systematic review [8].	This paper talks about the extensive amount of research that has been done on the use of machine learning for speech processing applications, especially speech recognition. It discusses the state-of-the results that deep learning approaches have given compared to others in various speech applications. This paper provides a thorough statistical analysis of the different studies that have been conducted since 2006 when deep learning first arose as a new area of machine learning for speech applications.	This paper focuses more on the theoretical research analysis of machine learning in speech. It fails to mention the upcoming resources and advancements that can be used to implement such techniques for creating a speech recognition system.
A survey of hybrid ANN/HMM models for automatic speech recognition [9].	This paper proposes several distinct structures as well as innovative training strategies. It examines various important hybrid models for ASR, bringing together ideas and techniques from heterogeneous and highly specialized literature. Efforts are focused on describing and referring to architectures and algorithms, grouping them into broad categories, and identifying their advantages and disadvantages.	The many components of the hybrid system employed in speech recognition are not discussed in depth in this paper.
A study on automatic speech recognition systems [10].	This paper aims to shed light on the basics of why and how speech has entered the world of HCI. It then discusses other related work that has been prominent in the field. This study then discusses the various stages employed in creating a speech recognition system and some notable industries where it is applied.	This paper gives only a short overview of the speech recognition system and does not shed light on the technicalities of the various stages of the speech recognition system. It also does not discuss the practical ways that a simple ASR system can be employed in today's world.
Wav2Letter: An End-to-End ConvNet-0based speech recognition system [11].	This research provides an end-to-end model for voice recognition that combines an acoustic model based on convolutional networks and graph decoding. It also presents the competitive results in word error rate on the Libri-speech corpus using MFCC features, as well as promising raw waveform findings.	This paper presents a detailed explanation only with regards to one particular model (Conv-Net) and focuses more on the mathematical implement- ation aspect rather than starting off with a baseline foundation

(Continued)

Table 1 (continued)

Paper	Content	Limitations
Exploring neural transducers for end-to-end speech recognition [12].	This study compares the CTC, RNN-Transducer, and attention-based models for end-to-end speech recognition empirically. It also analyses how the encoder architecture chosen influences the performance of the models.	Despite the fact that it compares three prominent models for the end-to-end ASR job at scale, these models differ in the simplicity of their training, which ultimately leads to the requirement of decoding with huge language models that were not thoroughly examined in the study.
Transformer-based acoustic modeling for hybrid speech recognition [13].	This paper proposes and evaluates transformer-based acoustic models (AMs) for hybrid speech recognition. It also discusses a variety of modelling options, including multiple positional embedding approaches and an iterated loss for training deep transformers.	The trials conducted in the study failed to indicate how much of transformer's higher performance is due to replacing recurrence with self-attention. Moreover, the other transformer modelling techniques which can be used to improve RNNs are not studied in detail.

The rest of the paper is divided into seven sections. The second section gives an overview of the factors depending on which ASR systems can be classified. Further, the third section provides the meat of the paper, summarizing the primary set of speech recognition techniques used. It includes some widely used feature extraction methods like Mel Frequency Cepstral Coefficients (MFCC) and Linear Prediction Coefficients (LPC) and their basic methodology. The following summarizes the building blocks of speech recognition systems, i.e., the models. Here, state-of-the-art models like Hidden Markov Model (HMM) and Deep Learning (DL) models have been highlighted, and some of the recent research work that has been conducted in this field have been compared [14]. The following section in the paper contains some of the best implementation tools for audio processing.

Furthermore, the fifth section briefly discusses the credible recent advancements in this field. This paper then discusses the most common industries where ASR systems are being extensively applied. Finally, the paper concludes its review by discussing the possible research gaps and future scope in speech recognition.

2 Speech Recognition System Classification

Speech is one of the most significant and primary modes of communication among human beings, making it a popular research domain [15]. However, humans in today's world are no longer limited to communication with each other. Instead, machines seem to have become an essential part of our

community. Expecting speech interfaces with machines, especially computers in native languages, seems to be the new face of technology. This interaction is done through interfaces; this area is called Human-Computer Interaction (HCI) [7]. The research in speech goes way back to the 20th century when computer scientists developed the first system to interpret and understand human speech [16]. Several years later, there is still scope for development and increased efficiency, which drives research efforts to make voice a viable human-computer interaction option [7].

In the computer system domain, speech recognition is described as a computer system's capacity to accept spoken words in an audio format, such as a wave or raw, and produce an understanding of the spoken information to identify it accurately [17].

ASR systems can be categorized depending on various factors, as mentioned ahead and shown in Fig. 1.

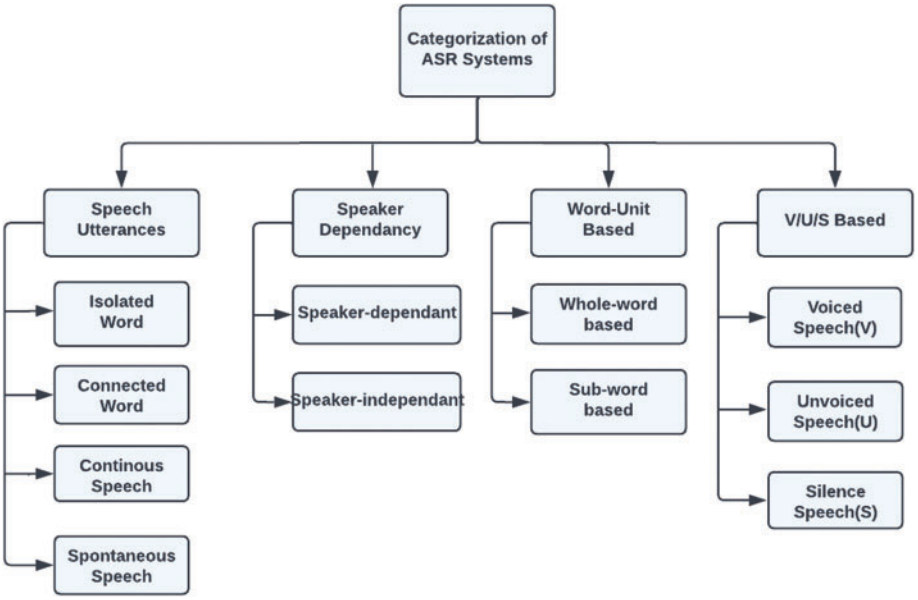

Figure 1: Categorization of ASR systems based on different factors

2.1 Speech Utterances

2.1.1 Isolated Word

This recognizer usually requires every utterance to have minimal or no noise on both sides of the sample window. It also allows only one utterance at a time, having "Listen/Not-Listen States", where the speaker must pause between the utterances [6].

2.1.2 Connected Word

Irrespective of their similarity with isolated words, separate words require a minimum amount of pausing between them, allowing them to almost run together [5].

2.1.3 Continuous Speech

Users can speak naturally through continuous speech recognizers while the computer figures out what they're saying. Because they need a unique way to identify utterance boundaries, recognizers with this type of speech capability are the most challenging to develop [6].

2.1.4 Spontaneous Speech

It is defined as a natural way of sound, unrehearsed speaking. Systems that recognize such utterances can handle various natural speech features [5].

2.2 Speaker Dependency

Speaker-dependent systems have been trained to recognize only one speaker, i.e., the person who will use the system. These systems can recognize words with a high degree of accuracy [6].

Speaker-independent systems are trained to respond to a word regardless of the speaker, which is challenging to achieve since they must respond to a wide range of speech patterns [6], inflections, and pronunciation of the target word, resulting in higher mistakes rates than speaker-dependent systems.

2.3 Word-Unit Based

A speech recognition system could be trained to special and unique keywords in their entirety, like "Ok Google" in Google Assistant and other such voice-command-systems [18]. These are useful in applications requiring only a small set of words to be recognized. However, this strategy, despite its simplicity, is not scalable. The recognizer's complexity and execution time increase as the recognized word dictionary grows [19].

On the other hand, sub-word unit-based speech recognition systems are used to identify sub-words like syllables/phonemes, followed by reconstruction of the words based on the same.

2.4 V/U/S Based

Speech recognized can be classified based on how the speech was produced. The vocal cords and the vocal tract produce phonemes, building blocks of any speech.

2.4.1 Voiced Speech (V)

When the vocal cords vibrate during the pronunciation of a phoneme, voiced signals are created. These signals are nearly periodic [20].

2.4.2 Unvoiced Speech (U)

On the other hand, unvoiced speech signals are created without using the vocal cords. Unvoiced speech signals are random noise-like, without any periodic nature [20].

2.4.3 Silence Speech (S)

Silence speech signals are created when there is no excitation supplied to the vocal tract, resulting in no speech output [20].

3 Speech Recognition Techniques

With the increase in the number of devices in the market, such as Siri in iPhone, Alexa from Amazon, speech recognition seems to have become quite influential in our everyday lives. With that being said, the baseline of speech recognition has always been allowing a machine to hear, to understand, and most importantly to act upon the basic information obtained, which makes the purpose of ASR to evaluate, extract, characterize, and recognize information about the spoken speech [5]. Hence, the importance of understanding and analyzing the techniques that go behind the complete

identification and perception of speech is quite evident. The speech recognition system combines three stages, as shown in Fig. 2.

- Feature extraction
- Modelling
- Performance evaluation

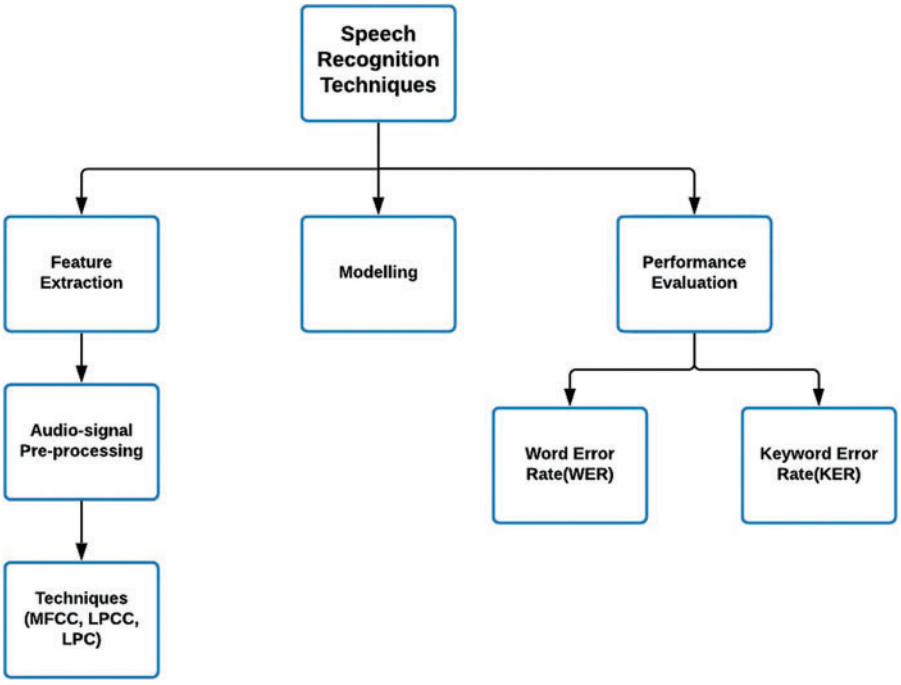

Figure 2: Primary subdivisions in speech recognition systems

3.1 Feature Extraction

The time-domain (amplitude *vs.* time) waveform of a speech signal carries all of the auditory information. However, to extract statistically significant information from the waveform, methods must be capable of condensing the information of every segment into a limited number of parameters, or characteristics, while keeping the signal's discriminating power [21]. A previous study has also found that when the number of input voice samples increases, the accuracy of speech recognition systems decreases [22]. The term feature extraction is introduced, which plays a crucial role in a speech processing system's accuracy. Feature extraction is a way of depicting a speech signal using a preset amount of signal components that are more discriminative and trustworthy. These features should characterize each segment so that other segments with similar features may be grouped by comparing their characteristics [23]. Feature extraction can be understood as one of the first steps in pre-processing (also called front-end signal-processing) of automatic speech recognition. Various approaches for extracting features from audio signals have been developed based on extensive research in mathematics, acoustics, and speech technology [21].

Feature Extraction goes hand in hand with model variable selection, which in turn contains another important sub section which can often make or break the performance. This is the model/feature selection algorithm that is chosen. Feature selection, even though never put on limelight has the

potential to drastically increase accuracy of a speech recognition system based on machine learning algorithms like random forest, SVM etc. This is because interaction between existing features of the same audio source can produce a number of redundant features and increase computational cost. The need and application of feature selection has been highlighted in papers that review speech emotion recognition in particular [24]. The different feature selection techniques often employed are correlation analysis [25], Linear discriminative analysis [26], fisher score [27,28] and Principal component analysis (PCA) [29].

Audio-signal Pre-processing: Before the feature extraction is carried out, pre-processing steps are first carried out on the raw audio signals, which are common to all techniques. These steps are shown in Fig. 3.

The four main pre-processing steps are:

- Analog-to-Digital Conversion: As sound is a mechanical wave, it is analog. However, it needs to be converted into digital signals to store audio in computers and used in machine learning systems. This conversion uses the ADC (Analog-to-Digital Converter), where an appropriate sampling frequency value is provided according to the application [32].
- Pre-emphasis: Raw speech signals can have quick energy reduction, leading to implementation issues in the practical world. Applying mathematical transform would result in different accuracy for different parts of the raw signal. Pre-emphasis, which emphasizes higher frequencies, is a typical pre-processing tool that compensates for the average spectral shape. It is accomplished by sending the signal in a finite impulse response (FIR) filter, commonly a first-order. FIRs filter [21–23].
- Framing: This process of partitioning the speech signal into several frames, each having some samples, is called framing. The reason for breaking the voice signal into short frames is that it is non-stationary, and its temporal features change rapidly. As a result, we assume that the speech signal will be stationary and that its features will not vary much inside the frame by using a small frame size. Frames allow us to consider speech signals periodic, which further helps perform Fourier analysis [33].
- Windowing: Windowing is then done on the framed signal, which helps in smoothing the endpoints by eliminating samples at both ends of a frame to generate a periodic function. Tapering the beginning conclusion of frame zero improves the sharpness of the harmonic and removes the discontinuity of signals. The two most famous categories are Hamming/Rectangular windows [21].

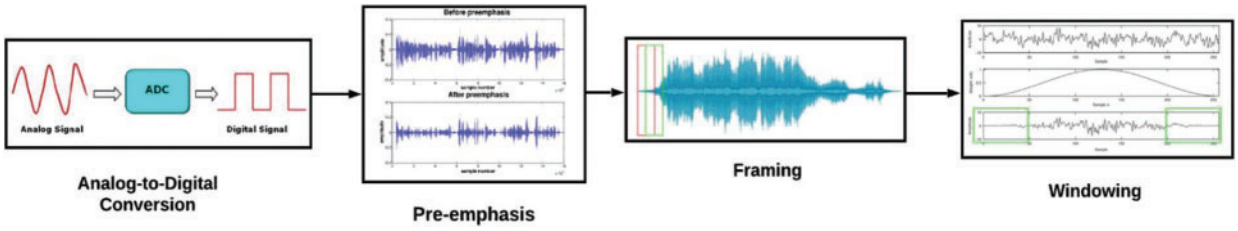

Figure 3: Basic Audio-signal pre-processing steps adopted from [30,31]

Apart from the above-mentioned basic and most common preprocessing methods, additional processing techniques are employed nowadays to refine the raw audio signal to a greater extent as shown in Fig. 4. They can be seen as:

- Voice Activity Detection: The major issue for the speech recognizer is obtaining or locating the endpoints of a signal in a speech. Incorrect endpoint detection will reduce the speech recognizer's performance. However, recognizing the endpoints of a speech utterance appears to be very simple, but has proven to be extremely challenging in practice in speech recognition systems. In a dynamic context, it is difficult to effectively model quiet and noise; removing voice and noise frames will make it easier to model speech. Furthermore, speech contains a large number of silent and noisy frames, which adds to the computational complexity. The removal of these frames reduces complexity while increasing accuracy. Voice activity detectors (VAD) aid in this process by simply splitting the speech signal into voiced or unvoiced segments, speech segments, and non-speech segments [36].
- Normalization: Feature normalization is a crucial step that is used to reduce speaker and recording variability without sacrificing feature discriminative strength [37]. The ability of features to generalize is improved by utilizing feature normalization. Normalization can be done at different levels in order to increase the efficiency of the ASR systems [38].
- Noise Reduction: The capacity to distinguish valuable sections of a voice signal from a stream of data might be critical during the early phases of an audio analysis system process. The term "ambient noise" refers to any signal that is not the signal being monitored. It is a type of interference or noise pollution. In fact, background noise is a key notion in determining noise levels in ASR systems. When training and testing data with different noise levels are used, the performance of speech recognition systems suffers dramatically. Hence, it is necessary to process noisy speech in order to minimize the effect of noise in the speech signals within the recognition system [39]. The most common strategy for reducing the effect of ambient noise on speech recognition is to employ a close-talk microphone [36].

Once appropriate steps in raw signal pre-processing are completed, feature extraction is next. This paper discusses two of the most common feature extraction techniques proven to be highly efficient and reliable in various speech recognition applications. These are MFCC and LPC.

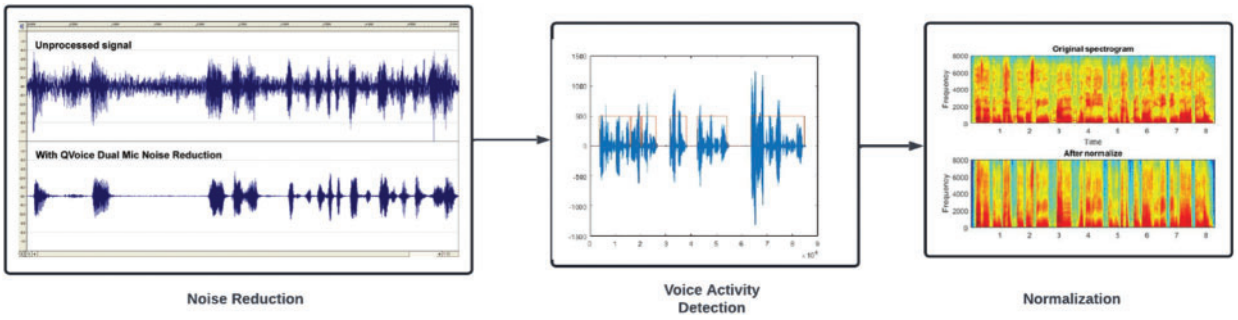

Figure 4: Additional Audio-signal pre-processing steps adopted from [34,35]

3.1.1 Mel Frequency Cepstral Coefficient (MFCC)

MFCC is the most popular feature extraction technique to the point where it is said to have become the standard feature extraction method for speech recognition [21]. Its popularity stems from the fact that it tries to mimic the human hearing system [40]. MFCC features are rooted in the recognized discrepancy of the human ear's critical bandwidths with frequency filters spaced linearly at low frequencies and logarithmically at high frequencies, which have been employed to keep the

phonetically crucial aspects of the speech signal [14–41]. By converting traditional frequency to Mel Scale, MFCC considers human perception for sensitivity at appropriate frequencies. The results in coefficients indicate features in the speech signal, such as power, pitch, and vocal tract shape [22].

The first three steps in the road map of MFCC comprise analog-to-digital conversion, windowing, and framing, which have already been mentioned above. Fig. 5 gives an overview of the flow of the undermentioned steps adopted in extracting features from the speech signal using MFCC:

- DFT (Discrete Fourier Transform): Discrete Fourier Transform converts the audio signals from the time domain into the frequency domain since analyzing speech signals in the frequency domain is much easier than in any other domain. Moreover, this particular domain helps obtain the most valuable features in speech analysis. After application, it results in the power spectrum of each frame [42].
- Mel Filter Bank: The step provides the disintegration of the signal with the help of a filter bank [43]. Human beings have the unusual tendency of differentiating between sounds only at a lower frequency range. On the contrary, machines resolve all sounds simultaneously, irrespective of their frequencies. Therefore, to enable machines to go hand in hand with humans, the Mel filter bank is brought into the picture, aiming to mimic the human ear perception of sound, thereby improving the model's performance later [44]. The mel scale is used to map the actual frequency to the frequency that human beings perceive. The formula for the mapping is [21–45].

$$melf = 2595 * \log[(101 + f700)] \tag{1}$$

where melf is Mel-frequency and f is the actual frequency

- Applying Log: The logarithmic function is applied to the output of the mel-filter bank since it has the similar property of producing a high gradient for an input having a less value and vice versa, which is required to simulate the way the human ear works. After using the Mel-Filter Banks and the logarithmic function, we are left with a spectrogram.
- DCT (Discrete Cosine Transform): The problem with the resulting spectrogram is that the Filter bank coefficients are discovered to be significantly correlated, giving us a reason to decorrelate them. As a result, DCT is used to convert the spectrum of Mel records into the time domain and the resulting list of integers or coefficients is referred to as the MFCCs (Mel Frequency Cepstral Coefficients) [46].

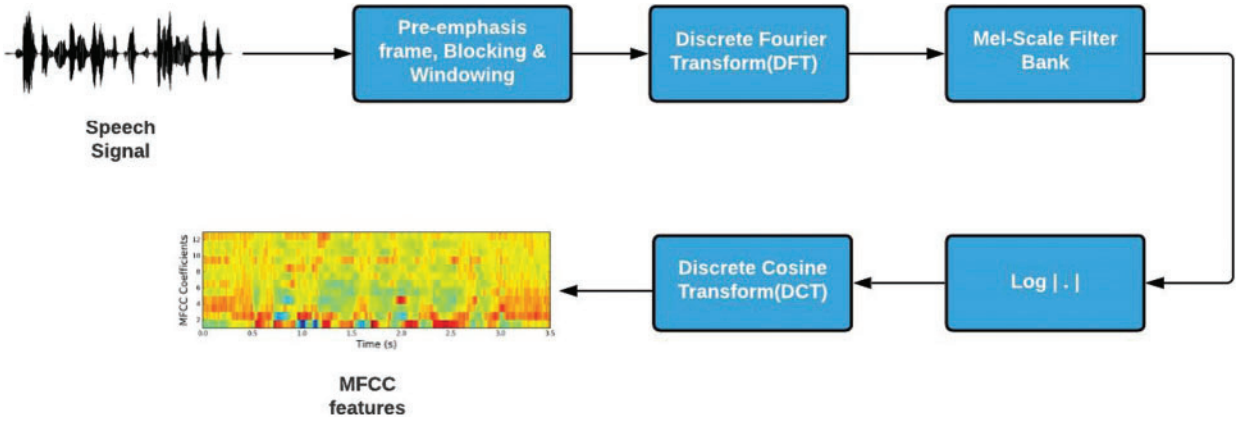

Figure 5: Workflow of feature extraction using MFCC, adopted from [21]

3.1.2 LPC (Linear-Prediction Coefficients) and LPCC (Linear Prediction-Cepstral Coefficients)

LPC is another essential and influential technique amongst other speech analysis methods. Its ability to encode high-quality speech at a reduced bit rate while simulating human vocal tract function has helped it gain market acceptability [47]. Linear prediction analysis of speech signals is based on the idea of forecasting each given speech sample at a specified time as a linear weighted aggregate of previous or prior samples [21]. While MFCC takes the nature of the speech into account when extracting features, LPC predicts future features based upon previous ones. It is commonly used in speech recognition systems with the primary purpose of extracting vocal tract features. It provides exceptionally accurate speech parameter estimates and is computationally intensive.

Methodology: LPC minimizes the sum of the squared discrepancies between the original and approximated speech signals during a finite time. Its methodology is such that every frame of the windowed signal is auto-correlated, and the highest auto-correlation value determines the LP analysis order [48]. The LPC analysis in which each frame of auto-correlations is turned into a set of parameters (LPC) comprises a unique set of predictor coefficients known as LPC coefficients [49]. However, these coefficients are rarely directly used in real-time speech recognition systems because of their significant variance. As a result, these coefficients are changed into cepstral coefficients, a more robust set of characteristics.

3.1.3 LPCC

Cepstral coefficients obtained via LPC are linear-prediction cepstral coefficients (LPCC) [21]. LPCC coefficients are derived by applying the Fourier transform on LPC's (logarithmic) magnitude spectrum. The cepstral analysis is frequently used in speech processing because of its ability to accurately describe speech waveforms and attributes with a limited number of features [21,50]. According to research, LPCC features provide lower error rates than LPC features on average. Other well-known feature extraction techniques, such as DWT, LSF, and PLP [21], have stood the perils of time and are widely utilized in ASR systems for a variety of reasons; nevertheless, no single methodology stands out as superior because they all depend on the specific application to be used. Fig. 6 below shows the workflow of feature extraction using LPC:

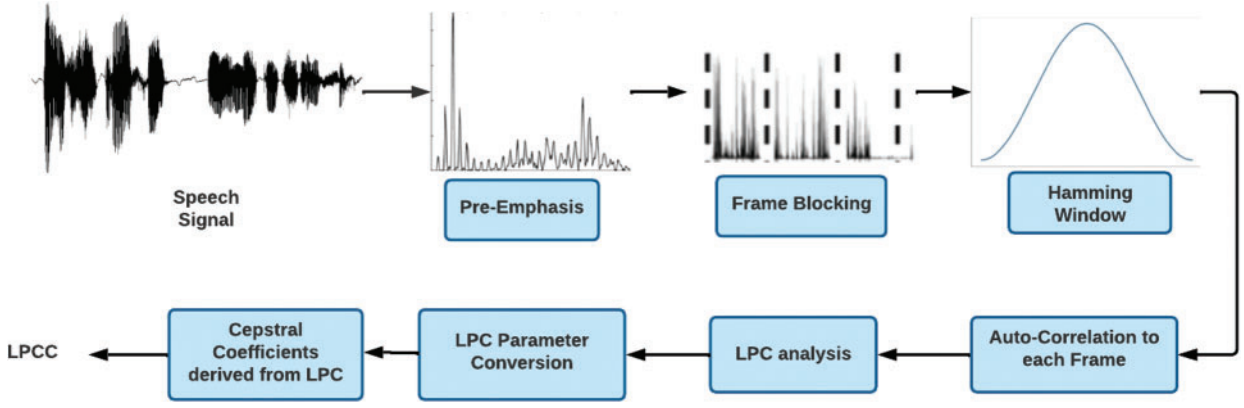

Figure 6: Workflow of feature extraction using LPCC

3.2 Modelling

After developing efficient feature vectors from the input audio signals, the next step is to move on to modeling or decoding these by choosing one of the various techniques available. Following are the three broad modeling approaches that are used as shown in Fig. 7:

- Acoustic phonetic approach
- Pattern recognition approach
- Deep learning approach

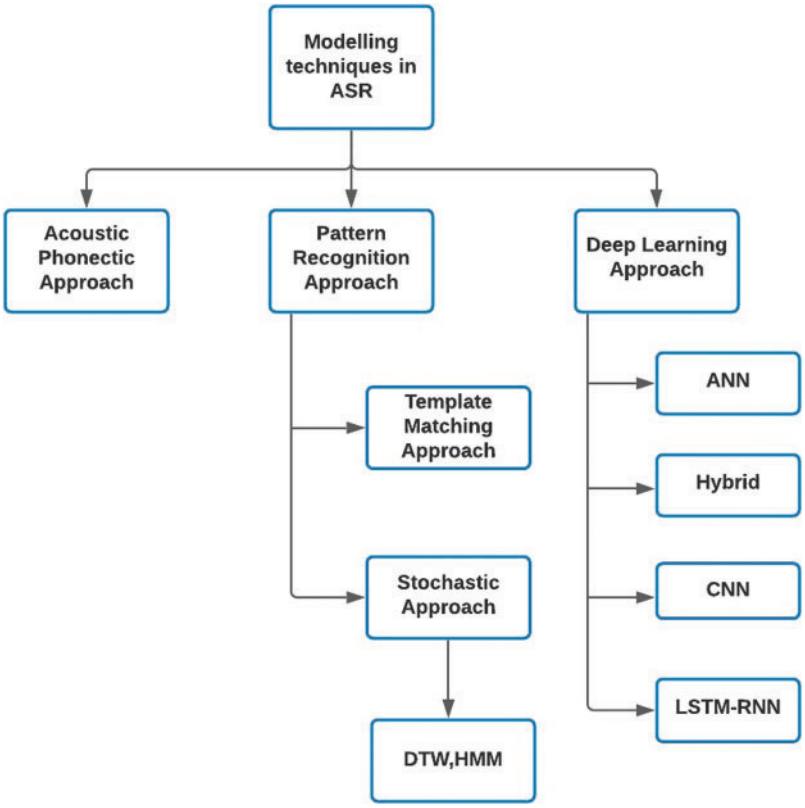

Figure 7: Classification of speech recognition techniques

3.2.1 Acoustic Phonetic Approach

The acoustic-phonetic approach is one of the oldest strategies for modeling ASR systems. This approach works by finding speech sounds and labeling them appropriately. The basis of this modeling technique is the belief that spoken language contains finite and exclusive phonemes that may be broadly described by a collection of acoustic qualities exhibited in the speech signal throughout time [51]. Although the acoustic qualities of phonetic units widely vary with speakers and surrounding sounds, the acoustic-phonetic method assumes that the principles underlying the variations are straightforward and may be mastered by a machine [5,52]. However, this particular concept is yet to be widely adopted in most applications in the commercial sector.

The Acoustic phonetic approach comprises mainly of two steps:

- The first is speech spectral analysis, which describes the various phonetic units' overall acoustic qualities. The speech is then segmented and labeled, resulting in a lattice of phonemes characterization of speech [53].
- Next, a string of words or a legitimate word is determined [6].

3.2.2 Pattern Recognition Approach

Pattern training and pattern comparison are the two fundamental aspects of the pattern recognition approach. The pattern-matching method was initially used in the 1900s [5]. The essential characteristic of this approach is that it employs an unambiguous framework of mathematics to create representations of constant speech patterns from a series of labeled training samples via a formal algorithm for training, allowing for reliable pattern comparison [5]. The basic idea underlying the pattern-comparison step (the practical element of the technique) is that the unknown speech (speech to be recognized) is directly compared to the patterns learned in the training stage [54]. Depending on the quality of the pattern matches, this comparison identifies or detects the identity of the unknown speech. In the last six decades, pattern-matching has been the most popular method for voice recognition [5].

Within this approach, there are two effective methods:

1. Template Matching Approach:

In this pattern recognition method, the underlying idea is simple to discover the best match; unknown speech is compared to a set of pre-recorded words [6]. A collection of prototypes depicting speech patterns, one for each candidate word called templates, are stored as references [55]. All of this is followed by recognizing the unknown speech for deciding on the best-matching pattern category. Even though it gives perfectly accurate word models, it has limitations. The predefined templates make these models non-flexible, while the extensive template preparation and matching make it prohibitively expensive, rendering it an impractical solution for growing vocabulary.

2. Stochastic approach:

On the other hand, this method uses probabilistic models to cope with ambiguous or inadequate speech data resulting from various sources, including confusing sounds, speaker variability, and so on, making it a particularly ideal approach for speech recognition. There are numerous approaches, such as HMM [56], DTW (Dynamic Time Warping) [57], and so on [6].

- Dynamic Time Warping (DTW): Dynamic time warping is a technique for determining how similar two series differ in speed and time [6]. DTW can analyze graphics, audio, videos, and other data converted into a linear representation. For example, similar walking patterns of two people, irrespective of speed, can be determined. Similarly, it can be applied to ASR to cope with different speaking speeds [5]. This method's primary goal is to generate a distance metric between two input time series [34]. The Euclidean distance between two points in vector space is used to calculate the similarity or dissimilarity of two-time series [58]. However, the DTW method finds matches with certain restrictions. These restrictions include the monotonicity of the mapping in the time dimension, high efficiency for only isolated word recognition, and the possibility of pathological results.
- Hidden Markov Models (HMM): Markov Models or HMM are one of the predominant and popular approaches that have been used under stochastic [5]. The sequences of states

that the model passes through are hidden and cannot be observed, hence the name Hidden Markov model. Its primary function consists of determining these states hidden from the observer and their respective probabilistic functions where one cannot determine the source or state from which an observation was produced. In reality, HMM is a statistical model in which the represented system is considered a Markov process with undetermined variables. The states are unknown in this case, but the variables controlled by the states are known. Hence, every state has a probable classification across the possible output tokens. Hence, the required information about the sequence of states is obtained from the sequence of tokens that the HMM generates [59].

HMM generates stochastic models using familiar words and analyses the likelihood that each of these models created the unknown utterance. This employs statistics postulates to organize feature vectors into a Markov matrix (chains) that contains state transition probabilities [60]. If each of our code words were to represent some state, the HMM would split the feature vector of the signal into a no. of states and find the chances of a signal to travel from one state to another state [61]. HMM's popularity stems from the fact that they can be trained automatically and are computationally feasible due to their solid mathematical foundation compared to the template-based approach discussed. Fig. 8 depicts a working HMM architecture.

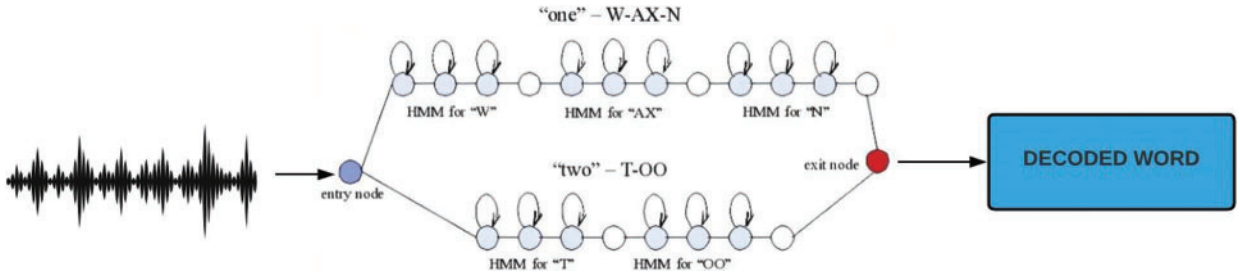

Figure 8: HMM for audio data processing [62]

Additionally, we also have the Viterbi Algorithm primarily used for prediction in Hidden Markov Models. Discovering the hidden state sequence in HMM that is most likely to have produced a given observation sequence is the most common problem in Hidden Markov Models. The solution to the problem is Viterbi Algorithm. Being an application of dynamic programming, it is widely used for estimation and detection problems in digital communications and signal processing [63]. It is used to detect signals in memory communication channels and to decode sequential error-control codes which in turn is used to improve the performance of digital communication systems, thereby explaining its primary use in speech and character recognition tasks where the speech signals or characters are modelled by hidden Markov models [64].

3.2.3 Deep Learning Approach

The deep learning approach has become the most commonly used approach in recent years. It owes its success to its ability to process the vast amount of information that ultimately automates the recognition procedure. Even though up until now, template-based approaches have been quite productive in the vast field of speech analysis, they provide little understanding of the working of human speech. The development of the artificial intelligence field in today's world has drawn researchers to look at the speech in a new data-driven methodology. This type of deep-learning-based system improvement has played a significant role in developing all effective solutions reported [65].

Under the deep learning approach, the different kinds of neural networks used today have brought up different angles of speech analysis. When alone or integrated with other methods, these neural networks bring about a credible increase in inefficiency. Some of these methods include:

1. Artificial Neural Networks (ANN)

Artificial neural networks (ANNs) are a computer-based model of biological neural networks [45–66], consisting of a directed graph with weighted and linked nodes. ANN's can handle the data residing on or near a non-linear model more effectively and learn better models of data [67]. Each network of ANN has multiple layers. Each layer is made up of weights, consisting of neurons or perceptrons. The activation functions trigger the neuron. The implementation of ANNs is divided into two parts: the training and testing phase. During the training phase, ANNs update weights continuously until the weights can fit the input data. After obtaining the desired weights, we can move on to the testing phase. The testing phase consists of new data not used by the ANNs known as the test set. The model's accuracy is measured using specific evaluation metrics according to the problem [68].

Architecture: Extracted feature vectors create a multilayer feedforward ANN architecture with supervised training. The backpropagation algorithm is used to train the ANN, which is effective in reducing recognition error rates [41,66,69]. A Backpropagation network consists of at least three layers: an input layer, one intermediate hidden layer, and an output layer. It is an approximate steepest descent algorithm. The performance index for backpropagation is mean square error.

2. Convolutional Neural Network (CNN)

CNN is a type of standard neural network. CNNs are predominantly used in image analysis but, with some appropriate changes, can be utilized in speech recognition as well [8]. The key advantage of a CNN-based model is its parameter efficiency [70]. Many proposed frameworks [71] have shown that using convolutional neural networks (CNNs) as opposed to other DNN approaches can result in significant error rate reduction due to its three main attributes: locality, weight sharing, and pooling, each of which has the capability of enhancing speech recognition performance. CNN's network differs from typical neurons in that it has a unique network topology in which convolutional and pooling layers are placed on top of each other as shown in Fig. 9. Each convolutional layer contains weights, and the pooling layer sub-samples the convolutional layer's outcome and decreases the underneath layer's DTR (data transfer rate). The weight sharing with properly chosen pooling schemes results from invariance in CNN. The feature maps obtained are passed through activation functions. The commonly used functions are ReLu, PReLu, and max out [72].

- Convolutional Layer: This layer comprises a collection of filters whose parameters must be learned. Each filter generates a neuron-based activation map. These maps of all filters are stacked over each other along dimensional depth to obtain the outcome of the layer [73].
- Pooling Layer: The pooling layer's primary goal is to reduce the feature map's size. A two-dimensional filter is slid across each map's channels during the pooling phase. The summarized features decrease the no. of parameters to learn, lowering the amount of computation in the network [74].

Before moving ahead with recurrent neural networks or types of recurrent neural network, it is necessary to be aware of the concept of Connectionist Temporal Classification function or commonly known as CTC. In the field of speech recognition, we often have a collection of audio clips and transcripts. Regrettably, we do not know how the characters in the transcript correspond to the audio. This makes training a speech recognizer more difficult than it appears at first. The simple approaches

are not available to us without this alignment. We could make a rule like "one-character equals ten inputs." However, because people's speech rates vary, this type of rule can always be broken. Another option is to manually align each character to its position in the audio. This works well in terms of modelling because we'd know the ground truth for each input time-step. However, for any reasonably sized dataset, this takes an inordinate amount of time. The CTC algorithm overcomes these challenges. Connectionist temporal classification (CTC) is a type of neural network output and associated scoring function used to train recurrent neural networks (RNNs) such as LSTM to solve sequence problems with variable timing. It is an approach for sequence labeling that uses a neural N-layer "encoder", which maps the input sequence to a sequence of hidden states, followed by a softmax to produce posterior probabilities of frame-level labels (referred to as "CTC labels") [75].

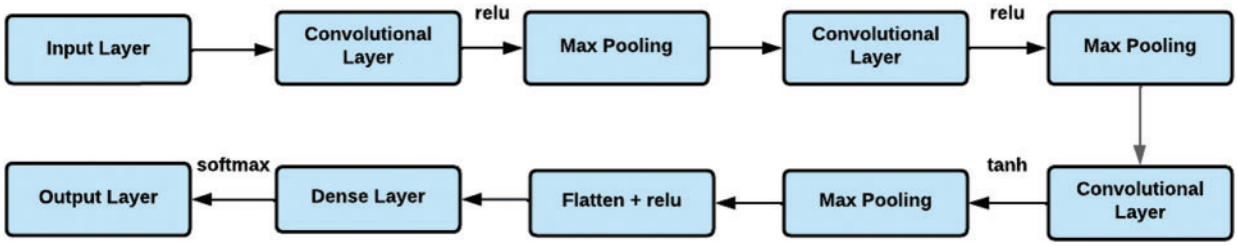

Figure 9: CNN architecture

CTC addresses two major issues in end-to-end ASR. To begin with, there is no longer any need to segment and align the speech data. CTC adds a blank label '-' that implies 'no output at this time.' It creates the path's intermediate structure based on the blank label. Some pathways can be swallowed into a final label sequence by deleting all repeated and blank labels. As a result, even in the absence of segmentation and alignment, CTC can map input sequence to the output sequence. Second, because CTC's output sequence is exactly what we expected, there is no need to develop extra modules to post-process it [76]. Moreover, this CTC operation computes the CTC loss between unaligned sequences. It computes the difference in loss between a continuous (unsegmented) time series and a target sequence and accomplishes this by summing the probability of possible input-target alignments, yielding a loss value that is differentiable with respect to each input node. Moreover, CTC loss is particularly designed for tasks where we need alignment between sequences, but where that alignment is difficult making it suitable for ASR Systems

3. Long Short-Term Memory-Recurrent Neural Network (LSTM-RNN)

Traditionally, standard feedforward neural networks did not have a mechanism to back information from the last layer to a previous layer. They could not handle speech recognition well. Hence, Recurrent Neural Networks (RNNs) came into the picture to account for the temporal relationships. However, even RNNS could not handle the long-term dependencies due to vanishing/gradient problems [77]. As a result, Long Short-Term Memory (LSTM) networks were developed, which are a subset of RNNs that account for long-term and short-term dependencies in speeches and control the flow of information by a particular unit called memory block [77,78]. These LSTM-RNN networks introduce the concept of gate functions, which indicate a neuron's activation value to transform or just pass through. This gating function causes the layer's inputs and outputs to be the same size [78]. Hence, quite interestingly, even though LSTM has a relatively simple architectural flow as shown in Fig. 10, its default habit is remembering information for lengthy periods, making it an appropriate tool for ASR systems [79]. LSTM-RNNs offer another benefit as they can make better use of parameters by distributing them over the space through multiple layers [80].

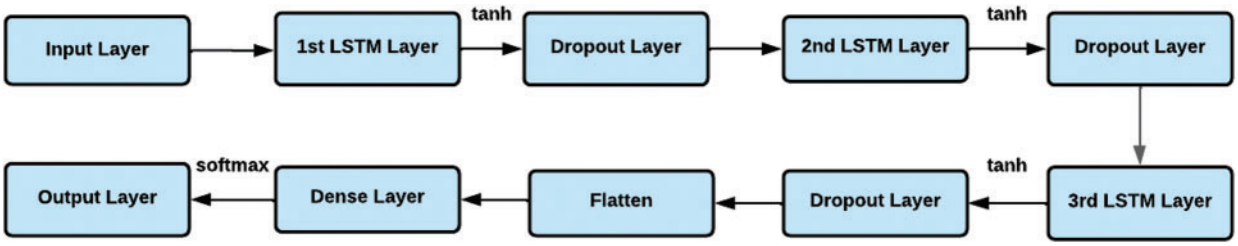

Figure 10: LSTM-RNN architecture

LSTM can be broadly divided into two categories:

- Unidirectional LSTM: By default, LSTM is unidirectional, i.e., it has a hidden state in a forward direction that processes the input from left to right by using the left context of the current input as shown in Fig. 11. This means it only gains access to a small bit of the right network context, leading to a somewhat lower recognition rate and low latency [78,81].
- Bidirectional LSTM (BLSTM): In bidirectional LSTM, on the other hand, input is provided bidirectionally, i.e., it has hidden states for the left and proper contexts, with a forward sequence for the left and a backward sequence for the right as shown in Fig. 12. Here, it uses different layers to handle data in both directions, forward and backward, thereby enhancing the identification rate [78].

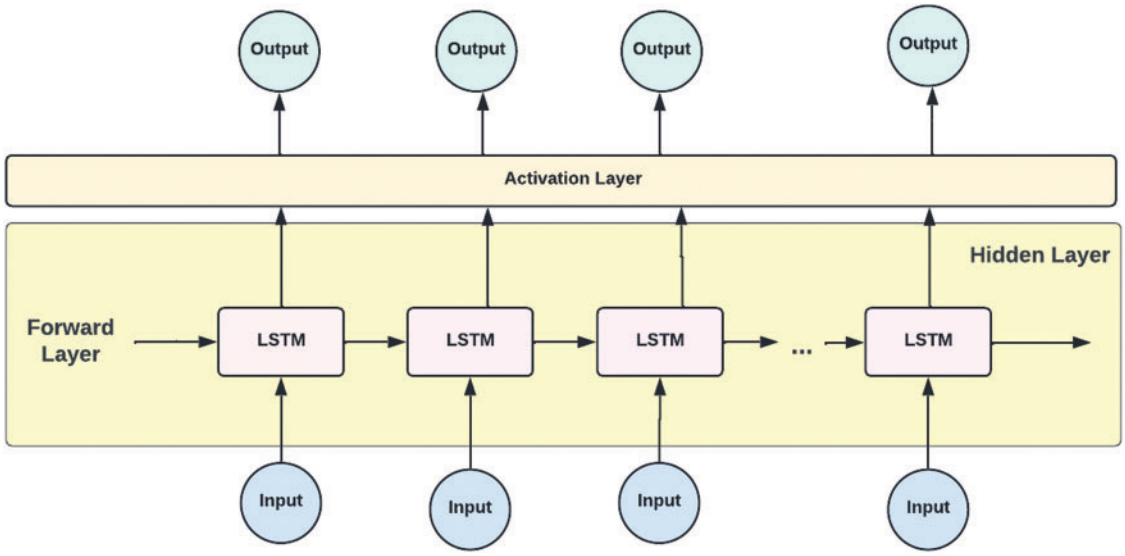

Figure 11: Unidirectional LSTM architecture

Most end-to-end ASR systems utilize a bi-directional Long Short-Term Memory (BLSTM) acoustic model due to its ability to capture the acoustic context from the entire utterance. However, BLSTM models do have high latency and cannot be used in streaming applications [82].

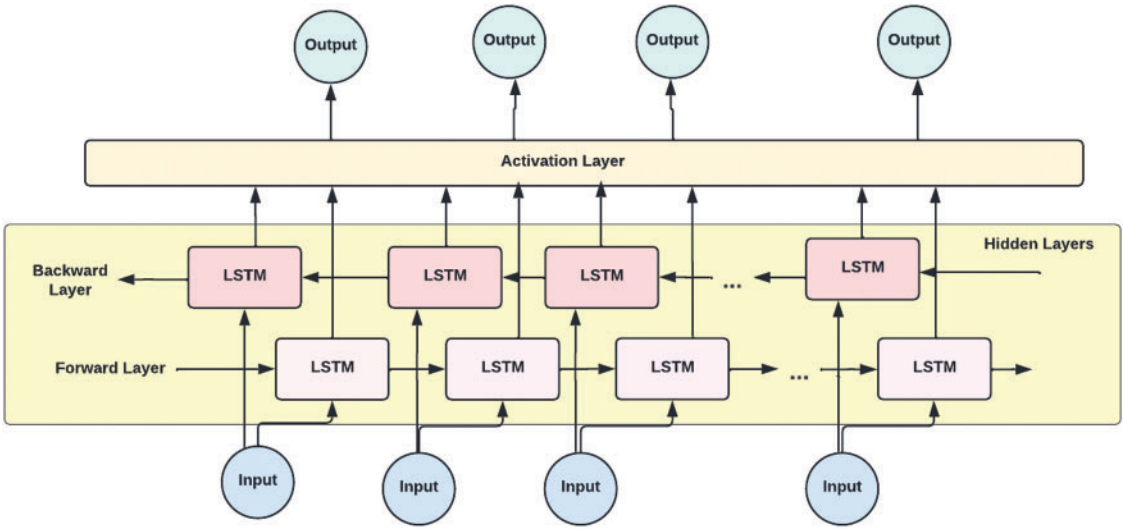

Figure 12: Bidirectional LSTM (BLSTM) architecture

According to the findings, bidirectional LSTM outperforms unidirectional LSTM and regular RNNs on the frame-wise phoneme classification test [78]. These findings imply that a bidirectional LSTM design is an appropriate architecture for speech processing in which context information is critical [77,78]. Tables 2 and 3 compare the accuracy of these LSTM models with various benchmark models.

Table 2: Comparative analysis of different experiments

Paper	Experiments performed by authors	Dataset	Feature extraction	Model	Model variable selection method	Metric used	Metric value (in %)
A neural attention model for speech command recognition [4]	de Andrade et al.	Google speech command dataset	MFCC	RNN	Attention mechanism	Accuracy	95.6
Convolutional neural networks for speech recognition [71]	Abdel-Hamid et al.	TIMIT phone recognition dataset	MFSC (mel-frequency spectral coefficients)	ANN	Correlation analysis	Word error rate (WER)	37.1
Convolutional neural networks for speech recognition [71]	Abdel-Hamid et al.	TIMIT phone recognition dataset	MFSC (mel-frequency spectral coefficients)	CNN	Correlation analysis	Word error rate (WER)	34.2

(Continued)

Table 2 (continued)

Paper	Experiments performed by authors	Dataset	Feature extraction	Model	Model variable selection method	Metric used	Metric value (in %)
Speech recognition for COVID-19 keywords using machine learning [89]	Amara et al.	COVID-19 keywords from recorded calls in Tunisian speech	FFT-Peak vectors	ANN	Grid search	Accuracy	97.0
Hidden Markov model-based speech emotion recognition [90]	Schuller et al.	Speech corpus	Pitch and energy contour	HMM	Discriminative feature analysis	Word Recognition Rate (WRR)	86.8
End-to-End mandarin speech recognition combining CNN and BLSTM [76]	Wang et al.	Mandarin speech corpus AISHELL-1	MFCC	CNN-BLSTM-CTC	None	Word error rate (WER)	19.2
Speech enhancement method based on LSTM neural network for speech recognition [79]	Liu et al.	XiaoMi corporation dataset	MFCC	LSTM	None	Word error rate (WER)	15.25

Table 3: Comparative analysis of different experiments with benchmark results

Area	Dataset	Model/Technique	Evaluation metric	Benchmark results
Neural attention model for speech command recognition	Google speech command dataset	Res8	Accuracy	94.1
Convolutional neural networks for speech recognition	TIMIT phone recognition dataset	DNN	WER	37.1
Speech recognition for COVID-19 Keywords using machine learning	COVID-19 keywords from recorded calls in Tunisian speech	SVM	Accuracy	93–97
Speech emotion recognition	Speech corpus	Human analysis	Accuracy	81.3
End-to-end speech recognition	Mandarin speech corpus AISHELL-1	CNN-input	Accuracy	20.68

(Continued)

Table 3 (continued)

Area	Dataset	Model/Technique	Evaluation metric	Benchmark results
Speech enhancement method for speech recognition	Xiaomi Corporation dataset	DNN	WER	17.53

As far as practical implementation is concerned, LSTM network topologies have outperformed the standard RNNs on learning Context-Free Language (CFL) and Context-Sensitive Language (CSL) as well as for sequence tagging and sequence prediction [77–79]. Moreover, RNNs and LSTMs, in particular, have performed exceptionally well in speech recognition tasks [77].

4. Hybrid Architectures

Another popular modeling technique for speech recognition today is hybrid architectures that combine two or more different modeling techniques so that their strengths can be put to use. Two of the hybrid network architectures that are most experimented with are:

- ANN-HMM Model: HMMs-based ASR systems are effective in many situations [44], but they have several significant drawbacks in real-world applications. The limitations were overcome using artificial neural networks (ANN); however, ANN failed to deal with lengthy timelines of speech inputs. Some academics began investigating a new research field in the early 1990s by integrating HMMs and ANNs into a sole, hybrid framework [9]. The main idea behind the hybrid architecture was to use the best of both approaches and develop an efficient solution [83].

  The architecture of the ANN-HMM Model as shown in Fig. 13 follows a sequential approach and relies on a probabilistic interpretation of the ANN outputs. Each ANN output unit is trained to compute a non-parametric estimate of an HMM state's posterior probability. This is a fundamental class of hybrid models that have a significant influence on a number of subsequent techniques. The relative ease of implementation, as well as a discriminative training criterion that may allow for higher recognition performance, are both strengths of this paradigm [84].

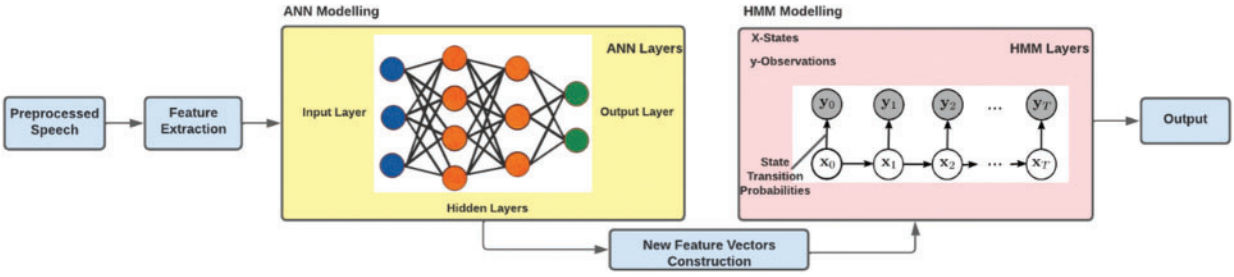

Figure 13: ANN-HMM architecture

- CNN-BLSTM Model: As discussed, earlier CNNs are a more advanced variant of DNNs that improve the word error rate by 4%–12% compared to DNNs on speech recognition tasks. This improvement is mainly due to spectral fluctuations and local correlations in speech signals that CNN handles better, making it better in speech recognition contexts. Bidirectional long

short-term memory (BLSTM) has recently been shown to improve recognition rates in acoustic modeling because it is capable of reinforcing higher-level representations of auditory data [77,78]. Combining both CNN and BLSTM to create a hybrid architecture has been gaining attention over the past few years. This hybrid architecture considers the audio signals' spatial and temporal properties required for a high recognition rate. The information from speech frames is captured by CNN, and BLSTM layers process this input in both directions. Initially, the convolutional layers present in CNN the feature map is reduced into much smaller size. As a result, there is no need to simulate locality or eliminate invariance any longer. Fully connected layers receive the output of BLSTM layers. These layers are well suited to producing higher-order feature representations that are easily distinguishable into various classes [78]. This hybrid architecture of CNN and BLSTM as shown in Fig. 14 has been used widely with other architectural elements, showing a relative decrease in WER over the general CNN and DNN architectural systems [76,77].

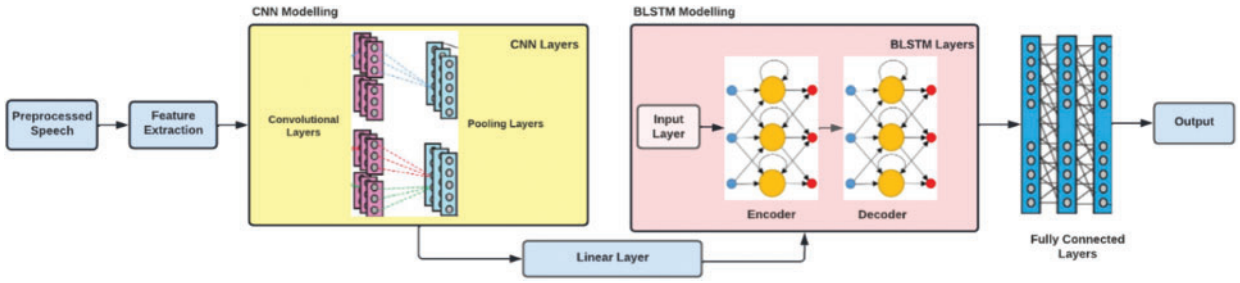

Figure 14: CNN-BLSTM architecture

3.3 Performance Evaluation Techniques

Accuracy and speed are necessary to determine the performance of speech recognition systems. Many factors influence vocalizations in speech recognition, such as pitch, pronunciation, articulation, roughness, volume, nasality, etc. Noise, echoes, and other factors can cause the speaker's speech to be distorted, making speech recognition difficult.

Apart from the factors influencing speech, several factors also influence the accuracy of the ASR system, like the size of vocabulary and speech confusability, speaker dependence and speaker independence, types of speech, language and task constraints, spontaneous speech, and uncertain environmental conditions. Hence, proper metrics are required for confidence measures of the system.

Various evaluation metrics can be used for this purpose, such as WER (Word error rate), SER (Sentence Error Rate), SWER (Semantic Word Error Rate), KER (Keyword Error Rate) to calculate the accuracy of ASR systems. Here, we have discussed the ones that show their best performance while working with speech recognition systems:

1) Word Error Rate (WER): The word error rate is a metric that is used to assess the performance accuracy of a voice recognition system. It captures the output transcript of an ASR [85]. When the length of the word of a recognized word varies from that of the reference word, the WER is calculated. The identified word sequence is aligned with the reference word sequence using dynamic word alignment. Dynamic Programming matching is performed to identify the correct words for each utterance and calculate WER. The formula to compute the word error rate of the system is given below:

$$WER = \frac{\mathrm{S} + \mathrm{I} + \mathrm{D}}{\mathrm{N}} \tag{2}$$

where S is the number of substitutions, D is the number of the deletions, I is the number of the insertions, and N is the number of words in the reference [86].

For evaluating the performance of systems, we can often use word recognition rate (WRR), which can be computed as given below.

$$WRR = 1 - WER = \frac{N - S - D - I}{N} = \frac{H - I}{N} \tag{3}$$

here, H is $N - (S + D)$, i.e., the number of correctly recognized words.

Speech recognition systems typically use WER as an evaluation metric. But many speech recognition applications also rely on identifying keywords and hence are more sensitive to keyword errors [87].

2) Keyword Error Rate (K.E.R.): Each Automatic speech recognition transcript is aligned with the reference transcript, and the KER is computed using the formula mentioned below.

$$KER = \left[\frac{\mathrm{F} + \mathrm{M}}{\mathrm{N}}\right] * 100 \tag{4}$$

where N is the number of keywords in the reference data, and F denotes the number of incorrectly identified keywords, M is the number of missed keywords. As recognition accuracy declines, the keyword mistake rate rises faster than the word error rate [88].

4 Implementation Tools

4.1 Libraries

One of today's essential tools to implement speech signal processing systems is the libraries explicitly created for audio analysis in different languages. A few of these are listed below:

- Librosa is used to implement some audio features (like MFCCs) using essential machine learning components (like clustering, etc.) [91].
- Pyaudioanalysis: A comprehensive library for audio signal processing in python. Pyaudioanalysis assists one in feature extraction, visualization of audio signals, training & applying audio classifiers, and segmentation of audio using supervised or unsupervised methods [92].
- Yaffe: A library written in python and used for audio feature extraction and primary audio input/output.
- Essentia: A library is written in C++ for audio analysis, audio feature extraction, and basic i/o [93].
- Bob: Written in python and C++, this library is used for signal processing and machine learning [94].
- Sinewave: A library used for sound analysis, including audio feature extraction and basic i/o. The library is an R package.
- MATLAB Audio Analysis: It is used for audio feature extraction, classification, and segmentation.

4.2 Datasets

Table 4 gives a comparative analysis of various datasets used to create speech recognition systems.

Table 4: Comparative analysis of datasets used in speech recognition

Name	Authors /Organization	Applications	Description	Dataset link
VoxCeleb	Nagrani et al.	Speaker Identification, Gender detection	A large-scale dataset consisting of almost 1,00,000 utterances, mostly gender-balanced [95]. This particular dataset serves as a compelling use case for the isolation and identification of individuals.	https://www.robots.ox.ac.uk/~vgg/data/voxceleb/
PF-STAR	Batliner et al.	Children speech identification	This dataset contains 1310 utterances in English from children aged between 4 to 14 years. It is mainly used in research on automatic recognition of children's speech [96].	http://www.thespeechark.com/pf-star-page.html
Urdu dataset	Latif et al.	Emotion Recognition for the Urdu language	It contains 400 utterances depicting four basic emotions: angry, happy, neutral, and emotion [97].	https://www.kaggle.com/datasets/bitlord/urdu-language-speech-dataset
The Berlin database of emotional speech (EMO-DB)	Burkhardt et al.	Emotion Recognition for speech uttered in the German language	This dataset features approximately 500 speech utterances spoken by actors in seven different emotions: pleased, furious, worried, afraid, bored, disgusted, and neutral [98]. Simple to use in regular communication, this dataset can interpret any emotion.	http://emodb.bilderbar.info/start.html
Librispeech corpus	Vassil Panayotov, Daniel Povey	Development of speech recognition systems	One of the most commonly used corpora is Librispeech, having about 1000 h of reading English speech derived primarily from the LibriVox project's read audiobooks [99].	https://www.openslr.org/12
TIMIT acoustic-phonetic continuous speech corpus	Garofolo et al.	Development of ASR systems	TIMIT features voices from about 630 speakers taken from broadband recordings, where each read ten phonetically rich sentences in a variety of American English dialects [100].	https://catalog.ldc.upenn.edu/LDC93S1
CHiME	Barker et al.	For noise-robust speech processing research	It is a noisy voice recognition challenge dataset typically less than 4 GB in size. Real (almost 9000 recordings of 4 speakers in 4 noisy settings), simulated (made by merging different environments across speech utterances), and clean voice recordings are all included in this collection [101]. Over the years, the authors have released various datasets, each with a unique setting.	https://archive.org/details/chime-home
Switchboard: Telephone speech corpus	Godfrey, Holliman, McDaniel et al.	Development of speech recognition systems	This corpus contains about 260 h of a well-separated collection of speech. Made up of approximately 2400 2-sided telephone conversations between 543 participants hailing from the USA.	https://www.isip.piconepress.com/projects/switchboard/

(Continued)

Table 4 (continued)

Name	Authors /Organization	Applications	Description	Dataset link
IEMOCAP: interactive emotional dyadic motion capture database	Busso et al.	Emotion Recognition	It contains audiovisual data from 10 actors of approximately 12 h, male and female, during improvised and scripted sessions. Five basic emotions are displayed mainly: happiness, anger, sadness, frustration, and neutrality [102].	https://sail.usc.edu/iemocap/iemocap_release.htm
RAVDESS (Ryerson audio-visual database of emotional speech and song)	Livingstone et al.	Speech emotion recognition	This is widely used. The dataset comprises actors and actresses (12 each), repeating the same lines. The emotions it can capture are standard like happy, calm, sad, etc., but each sets it apart at two intensity levels [103].	https://zenodo.org/record/1188976#.XrC7a5NKjOR
Speech accent archive	Steven H. Weinberger	Inclusivity of accents in speech recognition	Containing 2,140 English speech samples, this dataset is made by participants from 177 countries, each reading the same sentence.	https://www.kaggle.com/datasets/rtatman/speech-accent-archive/versions/1
Mozilla's common voice dataset	Mozilla	Development of speech recognition systems.	12 GB in size, Mozzila's common voice dataset contains 1000 s of voice samples taken mainly from blog entries, vintage movies, novels, and other public speech available, most of which are submitted by users [104].	https://commonvoice.mozilla.org/en/datasets
Google speech commands dataset	Warden et al.	To help train and evaluate keyword spotting systems	This dataset has sixty-five thousand clips, each lasting 1 s. Each clip contains one of the thirty different speech commands delivered by tens of thousands of people [105].	https://ai.googleblog.com/2017/08/launching-speech-commands-dataset.html
Fluent speech commands dataset	Fluent.ai	For end-to-end SLU.	With more than 30,000 utterances from approximately 100 speakers, this is an extensive and relatively new dataset. Each audio file in this collection comprises a single utterance for running intelligent appliances [106].	https://fluent.ai/fluent-speech-commands-a-dataset-for-spoken-language-understanding-research/

5 Recent Related Work/Advancements in ASR Systems

Voice recognition systems are not a far-fetched dream anymore as they used to be. Especially with the big giants of the world investing and competing every day to integrate speech technology into our daily lives. However, giants like Intel, Google, Nuance, and many small-scale experiments and research communities strive to develop 100% accurate ASRs. This section discusses several recent breakthroughs in voice recognition systems.

5.1 Automatic Speech Recognition Using Deep Learning

Deep neural networks are the latest technology that has contributed significantly to the development of speech recognition [107]. In 2015, Narayanan et al. [108] put forward a system using joint adaptive training to combine speech separation with acoustic modeling. The implementation was done with DNN, and the union was accomplished by adding hidden layers with fixed weights and the required network design. WERs were reduced by 10.9% (relative) compared to the noisy baseline when this approach was applied on the CHiME-2 [109] ASR task with Log Mel-spectral features as input. Xiong et al. [110], in 2016, first put forward conversational speech ASR from Microsoft, where they combined the numerous advancements in NN acoustic and language modeling to advance the accuracy of the Switchboard dataset. The system employs various CNN and RNN improvements for superior acoustic models, I-vector modeling, and lattice-free-MMI training. On the NIST 2000 Switchboard task, the best single system created used a ResNet architecture acoustic model and achieved a WER of 6.9%. The combined system further decreased the WER by 6.2 percent, improving the benchmark task results.

Further, in 2017, Xiong et al. [111] brought forward an updated version of the same Microsoft conversational speech recognition system. This newer version comprises techniques that are a part of NN-based acoustic and language modeling evolution. The Switchboard speech recognition task was improved by adding the CNN-BLSTM acoustic model to previous model architectures with character-based and dialog session-aware LSTM language models in rescoring. All of these were followed by system combination and confusion network rescoring. The result was a whopping 5.1% WER on the Switchboard dataset. In 2014, Chen et al. [112] proposed a Keyword Spotting application based on deep neural networks with a significantly small memory footprint, high amounts of precision, and low computational cost. The framework mentioned outperforms a regular HMM-based system in both clean and noisy situations. Keyword recognition results of 45 percent (clean) and 39 percent (babble noise) relative improvement over a competitive HMM model. This shows a more straightforward implementation, where a decoder is no longer needed, visibly reduced runtime computation, and a smaller and lighter model is achieved.

5.2 Speech Emotion Recognition (S.E.R.) Systems

Speech Emotion recognition is a relatively new and promising domain that is being extensively worked upon for the variety of new applications that it can cater to. These systems are primarily based on deep neural networks due to their high proficiency. Few are discussed ahead [113]. In 2019, Kwon [114] proposed an AI-assisted profound stride Convolutional Neural Network (CNN) architecture for speech emotion recognition. The framework uses a SoftMax classifier for emotion classification from speech. It is primarily based on learning salient and discriminative features from spectrograms of audio signals and hidden patterns in convolutional layers with special strides. The proposed technique is then evaluated on IEMOCAP and RAVDESS datasets, resulting in improved accuracy of 7.85% and 4.5%, respectively.

To recognize both (local & global) emotion-related features from speech, Zhao et al. [115] proposed the 2 CNN LSTM networks, a 1D CNN LSTM network and a 2D CNN LSTM network. The results demonstrated that both networks exhibit excellent performance on the selected databases, particularly the 2D CNN LSTM network, which beats the classic techniques of DBN and CNN. On the one hand, the 2D CNN LSTM network achieved 95.33 percent and 95.89 percent accuracies concerning speech emotion recognition on Berlin EmoDB's experiments for both types of speaker dependency, respectively, successfully showing better results as compared to traditional approaches, which achieved 91.6% and 92.9 percent accuracies. Similarly, on the IEMOCAP's experiments for both

types of speaker dependency, this network contributes 89.16 percent and 52.14% accuracies, which are much greater than the accuracies produced by DBN and CNN (73.78% and 40.02%, respectively).

5.3 Development of Tools for High-Performance Speech Recognition Systems

Park et al. [116] introduced a unique data augmentation tool for ASR named SpecAugment. This tool dramatically improves the performance of ASR that makes use of neural networks. It is a tool applied directly to feature inputs of a NN when a limited set of input data results in overfitting. It has obtained state-of-the-art results of 6.8% WER and 7.2% WER on the LibriSpeech 960 h [117] and Switchboard 300 h tasks end-to-end LAS networks by applying simple augmentation on the training dataset without the aid of a language model. Krishna et al. [69] have mentioned how electroencephalography (EEG) can be a promising technique to overcome performance loss due to background noise in ASRs. Several techniques have been employed in past research to make ASR systems robust and efficient in real-world scenarios by reducing background noise. However, this approach has revealed a significantly high recognition accuracy of 99.38% with background noise using EEG The model employed here is generally trained using 31 EEG channels + MFCC.

6 Applications of ASR Systems

Over the last few decades, speech recognition technology has evolved significantly, allowing a wide range of services and devices to become voice-enabled. There is a board spectrum of speech applications powered by ASR technology [118]. Earlier applications needed to be attended to, increasing production and maintenance costs. In a more practical view, addressing problems related to hand-free computing was increasing. These facts made automation the need of the hour, and speech is one of the primary modes of communication among human beings makes it an eligible entity for enhancing machine and human productivity. Exemplary user interfaces and promising dialogue models are the two aspects that form the foundation of a good ASR application. These aspects are developing rapidly with every passing day [119]. Therefore, speech recognition systems have worked their way from our cellphones to our houses, such as Amazon's Alexa, Apple's Siri, etc. At the same time, its application in different industries is becoming quite apparent. Some of these applications are mentioned in Fig. 15 and discussed below:

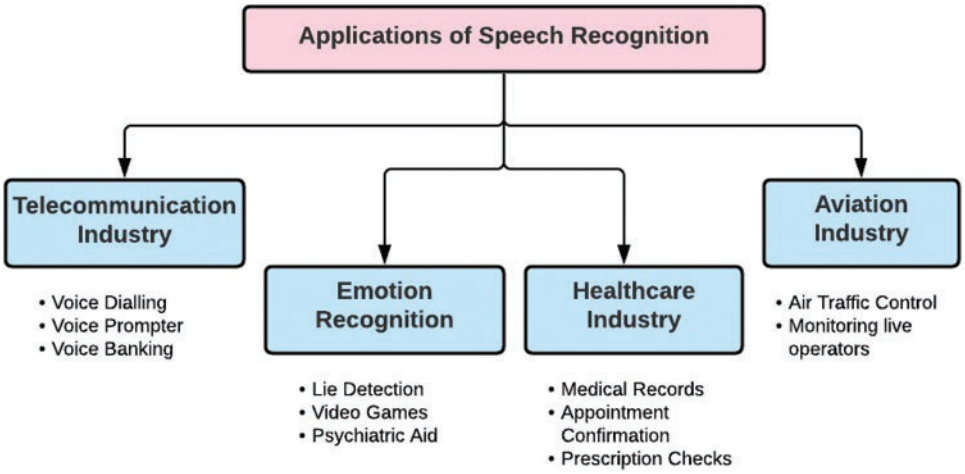

Figure 15: Different industries implementing speech recognition technology

6.1 Telecommunication Industry

The evolving network in the telecommunication field has shown the evident need for universal services. The telephone handset is the most crucial terminal device in this field and makes this industry dependent on speech recognition in command-and-control applications. Speech recognition has been introduced in telecommunication mainly to reduce costs and provide new revenue-generating services. Speech recognition provides a new way for telecom operators in the field of intelligent customer service [120]. Voice recognition call processing systems assist in automating operator services and voice dialing, whereas services that generate additional revenue include voice banking, voice prompter, and directory assistance call completion. These will eventually lead to reliable, powerful voice interfaces for all telecoms services, making them globally accessible [2].

6.2 Emotion Recognition

Humans have always found the speech to be the most natural medium of expressing themselves. Moreover, emotions convey considerable information about an individual, which opened up a new field, namely speech emotion recognition. In other words, it just came as an extension of this communication to computer applications by detecting embedded emotions through processing and classifying speech [121]. These speech emotion recognition methodologies have found a growing interest in the field of security systems concerned with the detection of lies, video games, psychiatric aid as well as in civil aviation to monitor the stress of aircraft pilots to help reduce the rate of a possible aircraft accidents [122].

6.3 Healthcare Industry

Recently, the healthcare sector has taken up a considerable chunk of speech recognition-based applications in the market [1]. The speech recognition technology produces effective results in the medical world since it provides instant feedback and prompt communication. From an application point of view, voice assistants help patients and managers retrieve information from medical records, confirm appointments and even check on prescriptions in detail. Doctors and patients can even benefit from the applications based on speech recognition technology through virtual communication between them. Moreover, ASR systems facilitate physically disabled people to command and control a machine, proving that speech recognition might soon take over a chunk of these patient management services.

6.4 Aviation Industry

Speech recognition in air traffic control has been a long-term research project in the research industry [123]. Simulating air traffic control, training, and monitoring live operators to improve safety are areas where work on integrating ASR with ATC is still ongoing [87]. Speech recognition can help balance and measure air traffic controller workload to limit the whole system's capacity. ASR can also find its use in transcribing controller pilot communications. Nevertheless, all of these applications are yet to turn into state-of-the-art possibilities. The way speech recognition evolves gives us a reason to believe in its efficacy in the air traffic control sector. Speech recognition technology can serve as a mode of communication between the aircraft captain and robotic co-pilots [124]. Apart from the industries above, speech recognition as technology has evolved in other sectors, such as the banking and marketing industry, media industry, various workplaces, intelligent homes, e-commerce industry, and many more [125].

7 Discussion and Recommendations

After a thorough review of all the aspects and experiments conducted concerning speech recognition, it could be widely observed from the works and research already conducted in the field that traditional machine learning models such as convolutional neural networks, recurrent neural networks and artificial neural networks have achieved remarkable accuracy in the field of speech, on the other hand hybrid models such as CNN-BLSTM are yet to reach a benchmark accuracy as far as implementations of ASR systems are concerned. As for ANN-HMM, even though both models standalone are strong and powerful models because of the strong mathematical structure of HMM [126] and preciseness of ANN, it is yet to reach a benchmark standard of accuracy. Crisp and clear audio dataset availability could be viewed as another concern, since a majority of publicly available datasets have background noises costing researchers a lot of computation time in cleaning the audio signals and producing refined ones. As far as applications of ASR systems are concerned, the majority of the visions are still on the plate owing to waiting for smooth pathways to be carved, thereby eliminating all the limitations posed by speech recognition in general.

Moreover, we observed that even though significant associated techniques and developments made in ASR systems, there is still room for improvement in various sectors, ranging from the accuracy of systems to their inclusivity. Human speech signals vary from speakers, speaking styles, content, and uncertain environmental noises [89]. Even though a significant amount of emphasis has been put on ASR, a lack of uniformity in speech systems is still a huge issue. Some issues are common to most systems today:

7.1 Variation in Accents

One major recognized issue that today's popular speech recognition systems fail to address is the wide variety of accents found in every language. The accent problem can be thought of as a domain adaptation. This bias is now gaining traction and leading to discussions on how technology needs to be more generalized and inclusive. The syllables and phonetics of the same word tend to fluctuate when spoken in various ways, making it even more difficult for the machine to process. Hence, one of the most common challenges with voice recognition is differences in pronunciation, accent, and intonation in general [127]. These disparities can be solved impeccably by using the exponentially growing amounts of data produced every day from people of different regions using different dialects in the training data for models. Today, many have to adapt to how they speak to interact with speech recognition technologies, especially those whose first language is not American English. This often can lead to an elevated identity crisis for marginalized communities like people with disabilities who rely heavily on voice recognition and speech-to-text tools [128]. Hence, it is a need for technology to no longer be disconnected from the complexity of human languages around the globe.

7.2 Children-Centric Focus

Another area that comes under the lack of inclusivity umbrella of speech recognition systems today is how they were never designed for use with children. With e-learning becoming a norm now, estimates say that children's screen time has surged by 60% [129]. ASR systems like Alexa, Siri, and Google should have been an asset to such growth, providing a more frictionless interaction with technology. However, children's speech recognition seems to be a tough nut to crack due to the acoustic and linguistic variability that a child's speech often brings to the table [130]. ASR for child speech is proven more challenging than that for adult speech, due to children's shorter vocal tracts, slower and more variable speaking rate and inaccurate articulation [131]. Children's voices, vocabulary, and behavior are significantly more complex. They make syntax, pronunciation, and grammar leaps that

speech recognition systems' natural language processing component must account for [129]. Speech recognition can be a valuable tool for children at home and in school. It has the immense potential to address critical gaps in children's reading and language development. This yet again implies that thriving ASR systems for young and ever-growing learners, who stand to benefit the most, is the most difficult. Speech recognition systems need to purposefully learn from the ways kids speak to account for and grasp the vastly different eccentricities of children's language [129].

7.3 Model Training

Deep learning is helpful in state-of-the-art ASR systems. However, overfitting is a severe problem with ANNs, which becomes apparent only when testing new data. Overfitting happens when a model learns the information and noise in the training data to the extent that it negatively impacts its performance on new data. This means that the model picks up on noise or random fluctuations in the training data and learns them as concepts. The problem is that these notions do not apply to new data, limiting the models' ability to generalize. Dropout training can be utilized to solve the problem. However, it appears to have been employed for word recognition in challenging situations. ANNs are not as successful in speech signals having long time sequences. Additionally, neural networks are better in accuracy and precision, but it needs high compositionality, longest processing time [132] and are not able to distinguish between very close sentences [133].

7.4 Noise

In regular discussion, we are frequently faced with the issue of communicating in less-than-ideal situations, such as when there is background noise. Human speech signals contribute a lot towards the generation of such "noise" or, as people might say, "noisy backgrounds." Moreover, Microphone and speaker configurations combine with the surrounding acoustic environment to create multiple challenges for traditional echo cancellation and noise suppression techniques [35]. In the field of voice recognition, removing the adverse effects of highly non-stationary environmental noise is an ongoing research topic [134]. An essential reason is a mismatch between the conditions in which a system is trained and used [135]. Many practical ways have been developed to solve this issue, including the ever-popular use of data-driven deep learning techniques. However, detecting and reducing the effects of unknown non-stationary noise remains a challenging task due to the unpredictable aspects of non-stationary noise [134].

Another problem with noise can be that it often results in severe data loss leading to more damage than planned. One of the ways that noise-robust speech recognition can be improved significantly is by creating and working with speech signals in more real-time environments. A few of such examples include using standard datasets in the noise section, such as-The CHiME challenge series [136] that provides speech signals in unique noisy environments like dinner parties, domestic environments, etc., the NOISEX, which is a database of recording of various noises like voice babble and factory noises as well as the SpEAR Database (Speech Enhancement and Assessment Resource) which provides noise-corrupted speech samples paired with clean speech references [137].

Further, ASR systems lack versatility due to their inability to automatically detect when they are not doing a good job of recognition and transfer control to human operators. Many niche developments in speech analysis, like the infamous hybrid modeling methodology [138] and transformer architectures [139], are yet to be applied extensively in real-life applications to know the advantages and limitations it can pose. Attempts to make ASR systems dynamic and noise-robust [140]; however, percent success is yet to be achieved.

8 Conclusion

Through the research presented in this paper, it is evident that several perspectives can be considered in developing a successful speech recognition system. Even though significant progress has been made in integrating speech signal processing in various systems, there is still immense scope to build an adequate ASR system. This paper has discussed and compared many techniques present in various stages of speech recognition systems in today's world and has found that the MFCC and deep learning techniques seem to be the backbone of the most accurate ASR systems. However, the known and unknown gap in speech signal analysis is still yet to be completely bridged. Speech recognition technology is being delved into extensively and has the scope to drive a future where algorithms recognize speech and the intention, behavior, and belief. With the increase in big data technology globally, there is increased scope of creating near-accurate and innovative voice systems.

Funding Statement: The Centre funds this research for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and Information Technology, University of Technology Sydney, Australia.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Latif, S., Qadir, J., Qayyum, A., Usama, M., Younis, S. (2021). Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering, 14,* 342–356. DOI 10.1109/RBME.2020.3006860.
2. Rabiner, L. R. (1997). Applications of speech recognition in the area of telecommunications. *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pp. 501–510. Santa Barbara, CA, USA. DOI 10.1109/ASRU.1997.659129.
3. Bandakkavanar, R. (2017). Speech recognition. https://krazytech.com/technical-papers/speech-recognition.
4. de Andrade, D. C., Leo, S., Viana, M. L. D. S., Bernkopf, C. (2018). A neural attention model for speech command recognition. http://dblp.uni-trier.de/db/journals/corr/corr1808.html#abs-1808-08929.
5. Gaikwad, S. K., Gawali, B. W., Yannawar, P. (2010). A review on speech recognition technique. *International Journal of Computer Applications, 10(3),* 16–24. DOI 10.5120/1462-1976.
6. Desai, N., Dhameliya, K., Desai, V. (2013). Feature extraction and classification techniques for speech recognition: A review. *International Journal of Emerging Technology and Advanced Engineering, 3(12),* 367–371.
7. Benkerzaz, S., Elmir, Y., Dennai, A. (2019). A study on automatic speech recognition. *Journal of Information Technology Review, 10(3),* 80–83. DOI 10.6025/jitr/2019/10/3/77-85.
8. Nassif, A. B., Shahin, L., Attili, L., Azzeh, M., Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE Access, 7,* 19143–19165. DOI 10.1109/ACCESS.2019.2896880.
9. Trentin, E., Gori, M. (2001). A survey of hybrid ANN/HMM models for automatic speech recognition. *Neurocomputing, 37,* 91–126. DOI 10.1016/S0925-2312(00)00308-8.
10. Ibrahim, H., Varol, A. (2020). A study on automatic speech recognition systems. *International Symposium on Digital Forensics and Security (ISDFS)*, pp. 1–5. Beirut, Lebanon. DOI 10.1109/ISDFS49300.2020.9116286.
11. Collobert, R., Puhrsch, C., Synnaeve, G. (2016). Wav2letter: An end-to-end convnet-based speech recognition system. arXiv preprint arXiv:1609.

12. Battenberg, E., Chen, J., Child, R., Coates, A., Li, Y. G. Y. (2017). Exploring neural transducers for end-to-end speech recognition. *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 206–213. Okinawa, Japan. DOI 10.1109/ASRU.2017.8268937.
13. Wang, Y., Mohamed, A., Le, D., Liu, C., Xiao, A. et al. (2020). Transformer-based acoustic modeling for hybrid speech recognition. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6874–6878. Barcelona, Spain.
14. Graves, A., Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. *ICML, 32,* 1764–1772.
15. Lee, W., Seong, J. J., Ozlu, B., Shim, B. S., Marakhimov, A. et al. (2021). Biosignal sensors and deep learning-based speech recognition: A review. *Sensors, 21(4),* 1399. DOI 10.3390/s21041399.
16. Malik, M., Malik, M. K., Mehmood, K., Makhdoom, I. (2021). Automatic speech recognition: A survey. *Multimedia Tools and Applications, 80(6),* 9411–9457. DOI 10.1007/s11042-020-10073-7.
17. Alharbi, S., Alrazgan, M., Alrashed, A., AlNomasi, T., Almojel, R. et al. (2021). Automatic speech recognition: Systematic literature review. *IEEE Access*. DOI 10.1109/ACCESS.2021.3112535.
18. Michaely, A. H., Zhang, X., Simko, G., Parada, C., Aleksic, P. (2017). Keyword spotting for Google assistant using contextual speech recognition. *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 272–278. Okinawa, Japan, IEEE.
19. Trivedi, N., Kumar, V., Sing, S., Ahuja, S., Chadha, R. (2011). Speech recognition by wavelet analysis. *International Journal of Computer Applications, 15(8),* 27–32. DOI 10.5120/1968-2635.
20. Rani, B. M. S., Rani, A. J., Ravi, T., Sree, M. D. (2014). Basic fundamental recognition of voiced unvoiced and silence region of a speech. *International Journal of Engineering and Advanced Technology (IJEAT), 2(2)*.
21. Alim, S. A., Rashid, N. K. A. (2018). *Some commonly used speech feature extraction algorithms*. London, UK: IntechOpen. DOI 10.5772/intechopen.80419.
22. Virkar, S., Kadam, A., Raut, N., Mallick, S., Tilekar, S. (2020). Proposed model of speech recognition using MFCC and D.N.N. *International Journal of Engineering Research and Technology (IJERT), 9(5),* DOI 10.17577/IJERTV9IS050421.
23. Shrawankar, U., Thakare, V. (2013). Techniques for feature extraction in speech recognition system: A comparative study. https://arxiv.org/abs/1305.1145.
24. Liu, Z. T., Wu, M., Cao, W. H., Mao, J. W., Xu, J. P. et al. (2018). Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing, 273,* 271–280. DOI 10.1016/j.neucom.2017.07.050.
25. Chadha, A. N., Zaveri, M. A., Sarvaiya, J. N. (2016). Optimal feature extraction and selection techniques for speech processing: A review. *2016 International Conference on Communication and Signal Processing (ICCSP)*, pp. 1669–1673. Melmaruvathur, India, IEEE.
26. Wu, C. H., Yan, G. L. (2004). Acoustic feature analysis and discriminative modeling of filled pauses for spontaneous speech recognition. In: *Real world speech processing*, pp. 17–30. Boston, MA: Springer.
27. Hegde, S., Achary, K. K., Shetty, S. (2015). Feature selection using fisher's ratio technique for automatic speech recognition. arXiv preprint arXiv:1505.03239.
28. Pacharne, M., Nayak, V. S. (2011). Feature selection using various hybrid algorithms for speech recognition. *International Conference on Computational Intelligence and Information Technology*, pp. 652–656. Berlin, Heidelberg, Springer.
29. Mitrović, D., Zeppelzauer, M., Eidenberger, H. (2009). On feature selection in environmental sound recognition. *2009 International Symposium ELMAR*, pp. 201–204. Zadar, Croatia, IEEE.
30. Velardo, V. (2021). How to extract audio features. https://github.com/musikalkemist/AudioSignalProcessingForML/blob/master/6-%20How%20to%20extract%20audio%20features/How%20to%20extract%20audio%20features%20.pdf.

31. Feng, L. (2014). *Speaker recognition* (*MS Thesis*). Technical University of Denmark, D.T.U.
32. Li, Q., Zhu, H., Qiao, F., Liu, X., Wei, Q. et al. (2018). Energy-efficient MFCC extraction architecture in mixed-signal domain for automatic speech recognition. *2018 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, pp. 1–3. Athens, Greece.
33. Hamid, O. K. (2018). Frame blocking and windowing speech signal. *Journal of Information, Communication, and Intelligence Systems (JICIS), 4,* 87–94.
34. Abusulaiman, N. S., Alhanjouri, M. A. (2017). Spoken arabic news classification based on speech features. *International Journal for Research in Applied Science and Engineering Technology, 5*. DOI 10.22214/ijraset.2017.8209.
35. Qsound Labs, Inc. (2011). https://www.qsound.com/products/qvoice.htm.
36. Ibrahim, Y. A., Odiketa, J. C., Ibiyemi, T. S. (2017). Preprocessing technique in automatic speech recognition for human computer interaction: An overview. *Annals of Computer Science and Information Systems, 15(1),* 186–191.
37. Kolokolov, A. S. (2002). Signal preprocessing for speech recognition. *Automation and Remote Control, 63(3),* 494–501. DOI 10.1023/A:1014714820229.
38. Akçay, M. B., Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication, 116,* 56–76. DOI 10.1016/j.specom.2019.12.001.
39. Garg, K., Jain, G. (2016). A comparative study of noise reduction techniques for automatic speech recognition systems. *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 2098–2103. Jaipur, India. DOI 10.1109/ICACCI.2016.7732361.
40. Bhatt, S., Jain, A., Dev, A. (2021). Feature extraction techniques with analysis of confusing words for speech recognition in the Hindi language. *Wireless Personal Communications, 118(4),* 3303–3333. DOI 10.1007/s11277-021-08181-0.
41. Chauhan, N., Isshiki, T., Li, D. (2019). Speaker recognition using L.P.C., MFCC, C.R.ZCR features with ANN and SVM classifier for large input database. *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, pp. 130–133. Singapore. DOI 10.1109/CCOMS.2019.8821751.
42. Kwek, L. C., Tan, A. W. C., Lim, H. S., Tan, C. H., Alaghbari, K. A. (2021). Sparse representation and reproduction of speech signals in complex Fourier basis. *International Journal of Speech Technology, 25,* 211–217.
43. Kiran, U. (2021). MFCC technique for speech recognition. https://www.analyticsvidhya.com/blog/2021/06/mfcc-technique-for-speech-recognition/.
44. López-Espejo, I., Tan, Z. H., Jensen, J. (2021). Exploring filterbank learning for keyword spotting. *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 331–335. Amsterdam, Netherlands, IEEE.
45. Martinez, J., Perez, H., Escamilla, E., Suzuki, M. M. (2012). Speaker recognition using mel frequency cepstral coefficients (MFCC) and Vector quantization (V.Q.) techniques. *CONIELECOMP 2012, 22nd International Conference on Electrical Communications and Computers*, pp. 248–251. DOI 10.1109/CONIELECOMP.2012.6189918.
46. Narkhede, A., Sen, N., Nemade, M. (2019). DCT application in speech recognition: A survey. *International Journal of Engineering and Techniques, 5(4),* 1–5.
47. Dave, N. (2013). Feature extraction methods LPC, PLP and MFCC in speech recognition. *International Journal for Advance Research in Engineering and Technology, 1,* 1–5.
48. Sanjaya, W. M., Anggraeni, D., Santika, I. P. (2018). Speech recognition using linear predictive coding (LPC) and adaptive neuro-fuzzy (ANFIS) to control 5 DoF Arm robot. *Journal of Physics: Conference Series, 1090(1)*. DOI 10.1088/1742-6596/1090/1/012046.

49. Paulraj, M. P., Sazali, Y., Nazri, A., Kumar, S. (2009). A speech recognition system for Malaysian English pronunciation using neural network. *Proceedings of the International Conference on Man-Machine Systems (ICoMMS)*, Batu Ferringhi, Penang, Malaysia.
50. Wu, Q. Z., Jou, I. C., Lee, S. Y. (1997). On-line signature verification using LPC cepstrum and neural networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 27(1),* 148–153. DOI 10.1109/3477.552197.
51. Arora, S. J., Singh, R. P. (2012). Automatic speech recognition: A review. *International Journal of Computer Applications, 60(9),* 1–11.
52. Satyanarayana, P. (2009). *Short segment analysis of speech for enhancement.* Institute of IIT Madras, Chennai, India.
53. Shinde, R. B., Pawar, V. P. (2012). A review on acoustic phonetic approach for Marathi speech recognition. *International Journal of Computer Applications, 59(2),* 40–44. DOI 10.5120/9523-3934.
54. Krishnan, K. S., Jerusha, K., Chugh, A. (2020). A review on speech recognition by machines. *International Journal of Engineering Research & Technology (IJERT), 9(7)*. DOI 10.17577/IJERTV9IS070016.
55. Gaudard, C., Aradilla, G., Bourlard, H. (2007). *Speech recognition based on template matching and phone posterior probabilities.* IDIAP.
56. Dixit, R. (2013). Speech recognition using stochastic approach: A review. *International Journal of Innovative Research in Science, Engineering and Technology, 2,* 1–8.
57. Pawar, R. V., Jalnekar, R. M., Chitode, J. S. (2018). Review of various stages in speaker recognition system, performance measures and recognition toolkits. *Analog Integrated Circuits and Signal Processing, 94(2),* 247–257. DOI 10.1007/s10470-017-1069-1.
58. Ismail, A., Abdlerazek, S., El-Henawy, I. M. (2020). Development of smart healthcare system based on speech recognition using support vector machine and dynamic time warping. *Sustainability, 12(6),* 2403. DOI 10.3390/su12062403.
59. Wang, D., Wang, X., Lv, S. (2019). An overview of end-to-end automatic speech recognition. *Symmetry, 11(8)*. DOI 10.3390/sym11081018.
60. Maseri, M., Mamat, M. (2020). Performance analysis of implemented MFCC and HMM-based speech recognition system. *2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, pp. 1–5. Kota Kinabalu, Malaysia.
61. Rupali, S., Sable, G. S. (2013). An overview of speech recognition using HMM. *International Journal of Computer Science and Mobile Computing, 2(6),* 233–238.
62. Pasquet, O. (2021). Search graph–HMM with phenomes. https://www.opasquet.fr/op-recognize/searchgraph/.
63. Li, Z., Zhu, Z., Guo, J. (2020). Research on HMM-based speech retrieval algorithm. *2020 IEEE International Conference on Artificial Intelligence and Electromechanical Automation (AIEA)*, pp. 122–126. Tianjin, China.
64. Lou, H. L. (1995). Implementing the viterbi algorithm. *IEEE Signal Processing Magazine, 12(5),* 42–52. DOI 10.1109/79.410439.
65. Seltzer, M. L., Yu, D., Wang, Y. (2013). An investigation of deep neural networks for noise robust speech recognition. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7398–7402. Vancouver, BC, Canada. DOI 10.1109/ICASSP.2013.6639100.
66. Nichie, A., Mills, G. A. (2013). Voice recognition using artificial neural networks and Gaussian mixture models. *International Journal of Engineering Science and Technology, 5(5),* 1120.
67. Dudhrejia, H., Shah, S. (2018). Speech recognition using neural networks. *International Journal of Engineering Research & Technology (IJERT), 7,* 1–7.
68. Hussain, S., Nazir, R., Javeed, U., Khan, S., Sofi, R. (2022). Speech recognition using artificial neural network. In: *Intelligent sustainable systems*, pp. 83–92. Singapore: Springer.

69. Krishna, G., Tran, C., Yu, J., Tewfik, A. H. (2019). Speech recognition with No speech or with noisy speech. *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1090–1094. Brighton, UK. DOI 10.1109/ICASSP.2019.8683453.
70. Han, W., Zhang, Z., Zhang, Y., Yu, J., Chiu, C. C. et al. (2020). Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. arXiv preprint arXiv:2005.03191.
71. Abdel-Hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G. et al. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(10),* 1533–1545. DOI 10.1109/TASLP.2014.2339736.
72. Zhang, W., Zhai, M., Huang, Z., Liu, C., Li, W. et al. (2019). Towards end-to-end speech recognition with deep multipath convolutional neural networks. In: *Intelligent robotics and applications*, pp. 332–341. DOI 10.1007/978-3-030-27529-7_29.
73. Kubanek, M., Bobulski, J., Kulawik, J. (2019). A method of speech coding for speech recognition using a convolutional neural network. *Symmetry, 11(9)*. DOI 10.3390/sym11091185.
74. Musaev, M., Khujayorov, I., Ochilov, M. (2019). Image approach to speech recognition on CNN. *Proceedings of the 2019 3rd International Symposium on Computer Science and Intelligent Control*, pp. 1–6. Amsterdam, Netherlands.
75. Krishna, K., Toshniwal, S., Livescu, K. (2018). Hierarchical multitask learning for ctc-based speech recognition. arXiv preprint arXiv:1807.06234.
76. Wang, D., Wang, X., Lv, S. (2019). End-to-end mandarin speech recognition combining CNN and BLSTM. *Symmetry, 11(5)*. DOI 10.3390/sym11050644.
77. Shewalkar, A., Nyavanandi, D., Ludwig, S. (2019). Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and G.R.U. *Journal of Artificial Intelligence and Soft Computing Research, 9,* 235–245. DOI 10.2478/jaiscr-2019-0006.
78. Passricha, V., Aggarwal, R. K. (2020). A hybrid of deep CNN and bidirectional LSTM for automatic speech recognition. *Journal of Intelligent Systems, 29(1),* 1261–1274. DOI 10.1515/jisys-2018-0372.
79. Liu, M., Wang, Y., Wang, J., Xie, X. (2018). Speech enhancement method based on LSTM neural network for speech recognition. *2018 14th IEEE International Conference on Signal Processing (ICSP)*, pp. 245–249. Beijing, China. DOI 10.1109/ICSP.2018.8652331.
80. Sak, H., Senior, A., Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. arXiv preprint arXiv:1402.1128.
81. Zen, H., Sak, H. (2015). Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4470–4474. Brisbane, Queensland, Australia.
82. Kurata, G., Audhkhasi, K. (2018). Improved knowledge distillation from bi-directional to uni-directional LSTM CTC for end-to-end speech recognition. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 411–417. Athens, Greece.
83. Gemello, R., Mana, F., Albesano, D. (2010). Hybrid HMM/neural network based speech recognition in loquendo ASR. http://www. loquendo. com/en/.
84. Tang, X. (2009). Hybrid hidden Markov model and artificial neural network for automatic speech recognition. *2009 Pacific-Asia Conference on Circuits, Communications and Systems*, pp. 682–685. Chengdu, China. DOI 10.1109/PACCS.2009.138.
85. Fish, R., Hu, Q., Boykin, S. (2003). Using audio quality to predict word error rate in an automatic speech recognition system. MITRE CORP BEDFORD MA.
86. Gamper, H., Emmanouilidou, D., Braun, S., Tashev, I. J. (2020). Predicting word error rate for reverberant speech. *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 491–495. Barcelona.

87. Nguyen, V. N., Holocene, H. (2015). Possibilities, challenges and the state of the art of automatic speech recognition in air traffic control. *International Journal of Computer and Information Engineering, 9(8)*. DOI 10.5281/zenodo.1108428.
88. Park, Y., Patwardhan, S., Visweswariah, K., Gates, S. C. (2008). An empirical analysis of word error rate and keyword error rate. *Interspeech,* 2070–2073. DOI 10.21437/Interspeech.2008-537.
89. Amara, W. B., Touihri, A., Hamza, S. (2020). Speech recognition for COVID-19 keywords using machine learning. *International Journal of Scientific Research in Computer Science and Engineering, 8(4),* 51–57.
90. Schuller, B., Rigoll, G., Lang, M. (2003). Hidden markov model-based speech emotion recognition. *2003 International Conference on Multimedia and Expo.* Shanghai, China. DOI 10.1109/ICME.2003.1220939.
91. Khurana, L., Chauhan, A., Naved, M., Singh, P. (2021). Speech recognition with deep learning. *Journal of Physics: Conference Series, 1854(1)*. DOI 10.1088/1742-6596/1854/1/012047.
92. Giannakopoulos, T. (2015). Pyaudioanalysis: An open-source python library for audio signal analysis. *PLoS One, 10(12)*. DOI 10.1371/journal.pone.0144610.
93. Bogdanov, D., Wack, N., Gómez, E., Sankalp, G., Herrera, P. et al. (2013). Essentia: An audio analysis library for music information retrieval. *14th Conference of the International Society for Music Information Retrieval (ISMIR)*, pp. 493–498. Curitiba, Brazil.
94. Anjos, A., El-Shafey, L., Wallace, R., Günther, M., McCool, C. et al. (2012). Bob: A free signal processing and machine learning toolbox for researchers. *Proceedings of the 20th ACM International Conference on Multimedia*, pp. 1449–1452. Nara, Japan.
95. Nagrani, A., Chung, J. S., Zisserman, A. (2017). Voxceleb: A large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612.
96. Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D. et al. (2005). The PF STAR children's speech corpus. *Interspeech,* 2761–2764. DOI 10.21437/Interspeech.2005.
97. Latif, S., Qayyum, A., Usman, M., Qadir, J. (2018). Cross lingual speech emotion recognition: Urdu *vs.* western languages. *2018 IEEE International Conference on Frontiers of Information Technology (FIT)*, pp. 88–93. Islamabad, Pakistan.
98. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B. (2005). A database of German emotional speech. *Interspeech, 5,* 1517–1520. DOI 10.21437/Interspeech.2005.
99. Kocabiyikoglu, A. C., Besacier, L., Kraif, O. (2018). Augmenting librispeech with French translations: A multimodal corpus for direct speech translation evaluation. arXiv preprint arXiv:1802.03142.
100. Garofolo, J. S. (1993). *TIMIT acoustic-phonetic continuous speech corpus.* Japan: Linguistic Data Consortium.
101. Barker, J., Marxer, R., Vincent, E., Watanabe, S. (2015). The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 504–511. Scottsdale, AZ, USA. DOI 10.1109/ASRU.2015.7404837.
102. Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation, 42(4),* 335–359. DOI 10.1007/s10579-008-9076-6.
103. Livingstone, S. R., Russo, F. A. (2018). The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS One, 13(5)*. DOI 10.1371/journal.pone.0196391.
104. Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M. et al. (2019). Common voice: A massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670.
105. Warden, P. (2018). Speech commands: A dataset for limited-vocabulary speech recognition. https://arxiv.org/abs/1804.03209.
106. Lugosch, L., Ravanelli, M., Ignoto, P., Tomar, V. S., Bengio, Y. (2019). Speech model pre-training for end-to-end spoken language understanding. arXiv preprint arXiv:1904.03670.

107. Deng, L., Li, J., Huang, J. T., Yao, K., Yu, D. et al. (2013). Recent advances in deep learning for speech research at Microsoft. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8604–8608. Vancouver, BC, Canada. DOI 10.1109/ICASSP.2013.6639345.
108. Narayanan, A., Wang, D. (2015). Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(1),* 92–101. DOI 10.1109/TASLP.2014.2372314.
109. Barker, J., Watanabe, S., Vincent, E., Trmal, J. (2018). The fifth "CHiME' speech separation and recognition challenge: Dataset, task and baselines. https://arxiv.org/abs/1803.10609.
110. Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X. et al. (2017). The microsoft 2016 conversational speech recognition system. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5255–5259. Calgary, AB, Canada. DOI 10.1109/ICASSP.2017.7953159.
111. Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X. et al. (2018). The microsoft 2017 conversational speech recognition system. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5934–5938. DOI 10.1109/ICASSP.2018.8461870.
112. Chen, G., Parada, C., Heigold, G. (2014). Small-footprint keyword spotting using deep neural networks. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4087–4091. Florence, Italy. DOI 10.1109/ICASSP.2014.6854370.
113. Lokhande, N. N., Nehe, N. S., Vikhe, P. S. (2012). Voice activity detection algorithm for speech recognition applications. *IJCA Proceedings on International Conference in Computational Intelligence (ICCIA2012)*, vol. 6, pp. 1–4. Levengipuram, Kanyakumari, India.
114. Kwon, S. A. (2019). A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors, 20(1),* 183. DOI 10.3390/s20010183.
115. Zhao, J., Mao, X., Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control, 47,* 312–323. DOI 10.1016/j.bspc.2018.08.035.
116. Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B. et al. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. https://arxiv.org/abs/1904.08779.
117. Panayotov, V., Chen, G., Povey, D., Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210. South Brisbane, QLD, Australia. DOI 10.1109/ICASSP.2015.7178964.
118. Wang, Y., Shi, Y., Zhang, F., Wu, C., Chan, J. et al. (2021). Transformer in action: A comparative study of transformer-based acoustic models for large scale speech recognition applications. *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6778–6782. Toronto, Canada.
119. Zeyer, A., Irie, K., Schlüter, R., Ney, H. (2018). Improved training of endto-end attention models for speech recognition. https://arxiv.org/abs/1805.03294.
120. Chen, H. (2019). *Success factors impacting artificial intelligence adoption: Perspective from the telecom industry in China* (*Doctoral Dissertation*). Old Dominion University.
121. Cohen, J. (2008). Embedded speech recognition applications in mobile phones: Status, trends, and challenges. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5352–5355. Las Vegas, NV, USA. DOI 10.1109/ICASSP.2008.4518869.
122. Lieskovská, E., Jakubec, M., Jarina, R., Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. *Electronics, 10(10)*. DOI 10.3390/electronics10101163.
123. McCallum, M. C., Campbell, J. L., Richman, J. B., Brown, J. L., Wiese, E. (2004). Speech recognition and in-vehicle telematics devices: Potential reductions in driver distraction. *International Journal of Speech Technology, 7(1),* 25–33. DOI 10.1023/B:IJST.0000004804.85334.35.
124. Wei, L., He, L., Liu, Y. (2020). Study of artificial intelligence flight co-pilot speech recognition technology. *2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*, pp. 681–685. Weihai, China.

125. Katore, M., Bachute, M. R. (2015). Speech based human machine interaction system for home automation. *2015 IEEE Bombay Section Symposium (IBSS)*, pp. 1–6. Mumbai. DOI 10.1109/IBSS.2015.745663.
126. Kardava, I., Antidze, J., Gulua, N. (2016). Solving the problem of the accents for speech recognition systems. *International Journal of Signal Processing Systems, 4(3),* 235–238. DOI 10.18178/ijsps.4.3.235-238.
127. Deshmukh, S. D., Bachute, M. R. (2013). Automatic speech and speaker recognition by mfcc, hmm and vector quantization. *International Journal of Engineering and Innovative Technology (IJEIT), 3(1),* 93–98.
128. Lloreda, C. L. (2020). Speech recognition tech is yet another example of bias. *Scientific American,* https://www.scientificamerican.com/article/speech-recognition-tech-is-yet-another-example-of-bias/.
129. Scanlon, P. (2020). Voice assistants don't work for kids: The problem with speech recognition in the classroom. *TechCrunch*. www.techcrunch.com/2020/09/09/voice-assistants-dont-work-for-kids-the-problem-with-speech-recognition-in-the-classroom/.
130. Dubagunta, S. P., Hande, K. S., Magimai.-Doss, M. (2019). Improving children speech recognition through feature learning from Raw speech signal. *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5736–5740. Brighton. DOI 10.1109/ICASSP.2019.8682826.
131. Feng, S., Kudina, O., Halpern, B. M., Scharenborg, O. (2021). Quantifying bias in automatic speech recognition. arXiv preprint arXiv:2103.15122.
132. Eljawad, L., Aljamaeen, R., Alsmadi, M. K., Marashdeh, I., Abouelmagd, H. et al. (2019). Arabic voice recognition using fuzzy logic and neural network. *International Journal of Applied Engineering Research, 14(3),* 651–662.
133. Al-Alaoui, M. A., Al-Kanj, L., Azar, J., Yaacoub, E. (2008). Speech recognition using artificial neural networks and hidden Markov models. *IEEE Multidisciplinary Engineering Education Magazine, 3(3),* 77–86.
134. Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A. E. D., Jin, W. et al. (2018). Deep learning for environmentally robust speech recognition: An overview of recent developments. *ACM Transactions on Intelligent Systems and Technology, 9(5),* 1–28. DOI 10.1145/3178115.
135. Haton, J. P. (1994). Problems and solutions for noisy speech recognition. *Journal de Physique IV Proceedings, EDP Sciences, 4(C5),* C5-439–C5-448. DOI 10.1051/jp4:1994592.
136. Barker, J. P., Marxer, R., Vincent, E., Watanabe, S. (2017). The CHiME challenges: Robust speech recognition in everyday environments. In: *New era for robust speech recognition*, pp. 327–344. Cham: Springer. DOI 10.1007/978-3-319-64680-0_14.
137. Andrew, V., Steeneken, H. J. M. (1993). Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication, 12*. DOI 10.1016/0167-6393(93)90095.
138. Shneiderman, B. (2000). The limits of speech recognition. *Communications of the ACM, 43(9),* 63–65. DOI 10.1145/348941.348990.
139. Viglino, T., Motlicek, P., Cernak, M. (2019). End-to-end accented speech recognition. *Interspeech,* 2140–2144. DOI 10.21437/Interspeech.2019.
140. Li, J., Deng, L., Gong, Y., Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(4),* 745–777. DOI 10.1109/TASLP.2014.2304637.