DESIGN AND DEVELOPMENT OF SPEECH RECOGNITION SYSTEM FOR MARATHI LANGUAGE

A Thesis

Submitted in partial fulfillment of the requirements for the

Award of the degree of

DOCTOR OF PHILOSOPHY
In

Computer Science and Engineering
By

Ravindra Parshuram Bachate
41800480

Supervised by

Dr. Ashok Sharma

Co-Supervised by

Dr. Amar Singh

LOVELY PROFESSIONAL UNIVERSITY
PUNJAB
2021

DECLARATION

I hereby affirm that the thesis entitled "Design and Development of Speech Recognition System for Marathi Language" submitted by me for the Degree of Doctor of Philosophy in Computer Science and Engineering is the result of my original and independent research work carried out under the guidance of Supervisor **Dr. Ashok Sharma**, and Co-supervisor **Dr. Amar Singh**, and it has not been submitted for to any university or institute for the award of any degree or diploma.

Place: LPU

Date: 13/01/2022

Signature of the Candidate

CERTIFICATE

This is certified that the thesis entitled — "Design and Development of Speech Recognition System for Marathi Language" which is being submitted by **Ravindra Parshuram Bachate** for the award of the degree of Doctor of Philosophy in Computer Science and Engineering, Lovely Professional University, Punjab, India is entirely based on the work carried out by him under our supervision and guidance. The work recorded embodies the original work of the candidate and has not been submitted to any university or institute for the award of any degree.

Place: LPU

Date: 13/01/2022

Dr. Ashok Sharma
(Supervisor)
Associate Professor
School of Computer Science and
Engineering, Lovely Professional University,
Phagwara, Punjab, India
Signature:

Dr. Amar Singh
(Co-Supervisor)
Associate Professor
School of Computer Applications
Lovely Professional University,
Phagwara, Punjab, India.
Signature:

ABSTRACT

In the beginning, humans used speech to communicate and exchange information with one another. The most obvious and natural method of human communication is communication through spoken words. Nowadays, there is considerable growth in the voice as an interface for robots, AI assistants, vehicles, mobile devices, etc. A human being usually uses a mouse, keyboard, or any other tangible interface to interact with a system or product. Designing the inclusive voice as an interface for the products needs attention to develop regional languages Speech Recognition System (SRS). This research focuses on the development of the Marathi Language Speech Recognition System which is spoken in Maharashtra, Goa, some parts of Karnataka, and Madhya Pradesh by considering various dialects, different speakers, and conditions. This research has proposed four Marathi SRS methodologies such as the MFCC-SF feature extraction approach, RCBO-DBN, GOA-RNN, and PB3C-LSTM pattern recognition approaches.

This research followed the approach of following stages - preprocessing, feature extraction and selection, and pattern recognition. The speech corpus for this research is collected from the Linguistic Department, Government of India. The Smoothing and medium filtering techniques were used in pre-processing of the speech signals. After that, the MFCC-SF approach was used for the feature extraction, which gave better results than MFCC and spectral features if used individually.

During this study, we investigated several techniques for building a voice recognition system for Marathi. To evaluate the performances of these techniques using various metrics, including accuracy, sensitivity, FPR, FNR, FDR and MCC, we examined classification methods, such as Support Vector Machine (SVM), k-Nearest Neighbor (KNN), Artificial Neural Network (ANN) and Deep Belief Network (DBN). DBN performance of the recognition compared to four other pattern recognition approaches was good.DBN gives about 81% accuracy.

The second approach developed for t Marathi SRS is RNN-based GOA. This approach has consisted of three stages such as pre-processing, feature extraction, and classification. The

input signals were pre-processed, which was further subjected to the feature extraction stage. Here, the MFCC and spectral-based features were extracted for the proposed speech recognition model. These features were classified using optimized RNN, where the number of hidden neurons was optimized using GOA. Finally, the proposed model has efficiently attained recognized speech. From the experimental results, the accuracy of the proposed GOA-RNN model was 5.2%, 1.16%, and 0.86% progressed than RNN, LSTM, and Res-CNN, respectively for test case 1. Therefore, from the experimental results, the WER of the proposed GOA-RNN model was 3.84%, 1.06%, and 0.79% improved than RNN, LSTM, and CNN, respectively, for speech corpus one, and it has similar results with remaining speech corpus.

P3BC-LSTM is a third approach proposed for developing Marathi SRS The feature extraction was carried out by MFCC and spectral-based features. Further, these features were significantly selected using PCA, which were forwarded to the classification stage using P3BC-LSTM. The P3BC-LSTM was proposed by optimizing the number of hidden neurons and weights using the P3BC algorithm that intended to get the recognized speech signals. Therefore, from the experimental results, the Word Accuracy Rate (WAR) of the proposed speech recognition model using PB3C-LSTM was 1.34% and 3.34% increased than LSTM and BB-BC-LSTM, respectively, while considering the Speech corpus one. The WER of the proposed P3BC-LSTM model has attained 4% and 6% less WER, and 4.45% and 3.25% less SER than LSTM and BB-BC-LSTM, respectively, for speech corpus one, and it has similar results with remaining speech corpus.

ACKNOWLEDGEMENT

I want to express all praises to God and my spiritual master Sadhguru for completing this research work and thesis. Many people's support and cooperation made the present work completed. I would like to express my respect and a deep sense of gratitude to my Supervisor, Dr. Ashok Sharma, and Co-Supervisor, Dr. Amar Singh, for his guidance, encouragement, and support in every stage of my research work. Their enthusiasm, inspiration, and encouragement helped me carry out the research and complete the work in various ways. I would like to express my gratitude to my research colleagues for their consistent support and encouragement. My completion of the thesis could not have been accomplished without the support of my parents, wife, and friends.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Introduction

Human-spoken languages are referred to as Natural Languages. Several sites or resources generate data every day in natural languages[1]. For example, Facebook, Twitter, Instagram, Blogs, etc., generate extensive data that are difficult to handle with conventional data processing tools. Nowadays, the world is driven by data, and we must focus on approaches that increase human living standards. Natural Language Processing (NLP) is crucial since we deal with user sentiments on different platforms. Many studies have been on NLP, but it is confined exclusively to select commonly used languages, e.g., English, Chinese, etc. When we think about India, almost 70 percent of the indigenous live in rural areas where English is not well understood. We must work with local languages such as Marathi, Hindi, Punjabi and improve their lives using NLP. There are 22 official languages spoken in India and more than 1000 dialects (Huang and Deng, 2010). These linguistic areas and dialects belong to many language families, including Indo-Aryan, Dravidian, Austroasiatic, Tibeto-Burman, etc. [2].

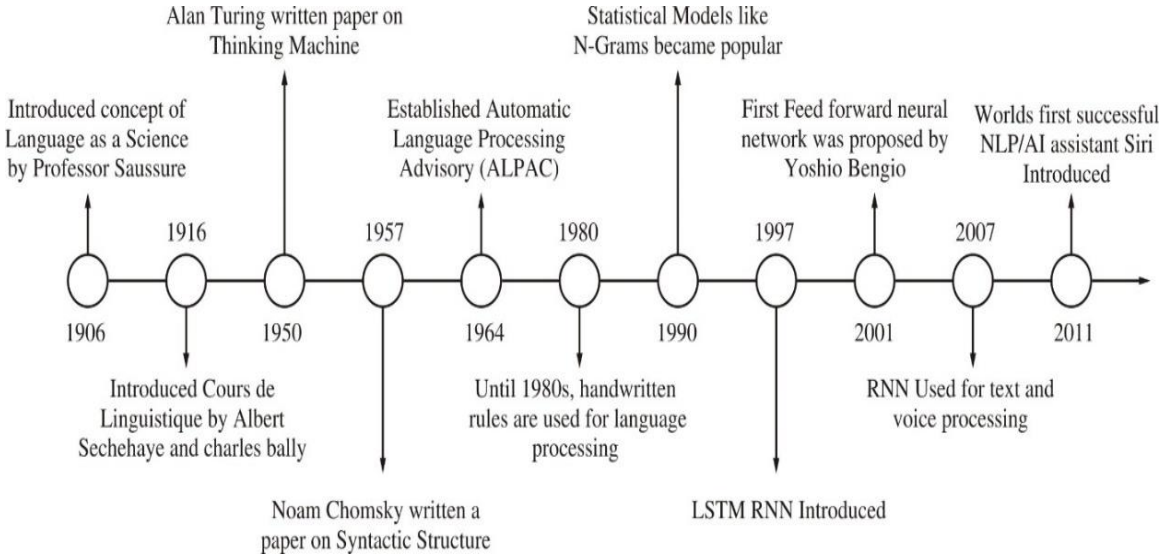


Fig. 1.1: History of Natural Language Processing

There are two aspects of the natural language - Natural language processing and natural language understanding. The Natural Language processing process addresses many tasks such as Named Entity Recognition, POS, classification of texts, means of data, conference resolution, and machine translation.

At the same time, natural language understanding deals with relation extraction, semantic parsing, questionnaire, sentiment analysis, synthesis, paraphrasing, and interaction with natural languages and dialogue bots.

The journey of natural language processing in the previous 10 decades can assist us to comprehend the advancement of natural language processing. Professor Saussure spoke on natural language as a science in 1906, and others have been investigating it ever since. From 1906 through 1911, he studied and researched natural language. His colleagues Albert and Cherles bally wrote in 1916 the article "Cours de Linguistique." Alan Turing wrote breakthrough work on Thinking Machine in 1950. A book on the syntactic structure was produced in 1957 by Noam Chomsky. In 1964, in partnership with the National NLP and AI Research Committee, the US formed the "Automatic Language Processing Advisory Committee (ALPAC)." By the 1980s, it was difficult for humans to establish handwritten rules for the processing of natural language, statistical models were established in the 1990s to address this challenge. Due to its employment in recognizing and monitoring linguistic data clumps, N-gram became popular [3].

Apple released the first NLP/AI assistant in the world, Siri. Later more company apps such as Google Assistant, Amazon's Alexa, Microsoft's Cortona, etc., were launched. In addition to the important contribution of applied machine learning to expanding the scope of application and the efficiency of natural language processing. Speech Recognition and Understanding, a natural language processing sub-domain, are utilized extensively in applications like Siri, Alexa, and more.

"Speech is a multi-component signal with varying time, frequency and amplitude”[4]. Human beings engage with one another or communicate using Speech. In comparison with non-Indian languages, there has been a rare study done on Indian languages. Many researchers throughout the globe strive to design a new human-computer interface system

with optimum precision. Speech may be an essential way of computer interface. Speech Recognition is the technique to process a voice in words or text form. The voice of speech data uttered by people is transformed into electronic signals using speech recognition technology. Then these signals are translated into a pattern of coding, and desired significance is achieved. Speech recognition is essentially the technique by which the specific speaker may be identified with the use of speech wave information. Researchers have been working on language recognition throughout the previous sixty years. Now ASR identifies an app that needs human-machine interaction and can talk and speak the language in their mother tongues [5].

Auditory System for people

It is vital to understand the functioning of the human auditory system to create the speech-based interaction system. The thought is first to come into the human's mind, and then it is translated into a speech. The notion is subsequently turned into words and sentences according to the language's grammatical norms. The brain produces electrical impulses that travel along the nerves at the physiological level of communication. These electrical impulses stimulate the vocal and vocal cord muscles. These vocal tract movements and the cord led to variations in the pressure in the vocal tract, particularly on the lips.

Mechanism of Speech Production

The interplay of the diaphragm, lung, throat, mouth, and nasal cavity causes human Speech. Phonation, resonance, and articulation are the mechanisms that regulate the creation of Speech. Phonation is the process employing vocal folds or vocal cords that convert air pressure into sound. Resonance is the mechanism through which resonances in the vocal tract stress-specific frequencies. The process of articulation is to change the resonance of the vocal tube to generate distinctive sounds.

Air enters the lungs via the usual breathing system. The tension vocal cord inside the larynx is induced to vibrate by airflow, as the air is released from the lungs via the trachea. The airflow is cut into very regular pulses that are often modified by travelling via the throat, the cavity of the mouth, and a nasal cavity. Different noises are created depending on the location of different articulators. Fig. 1.2 shows a simplified picture of the whole

physiological processes to create speech sound. The lungs and accompanying muscles operate as an air supply for the vocal System to be excited. The muscular strength pushes air through the trachea and bronchi from the lungs.

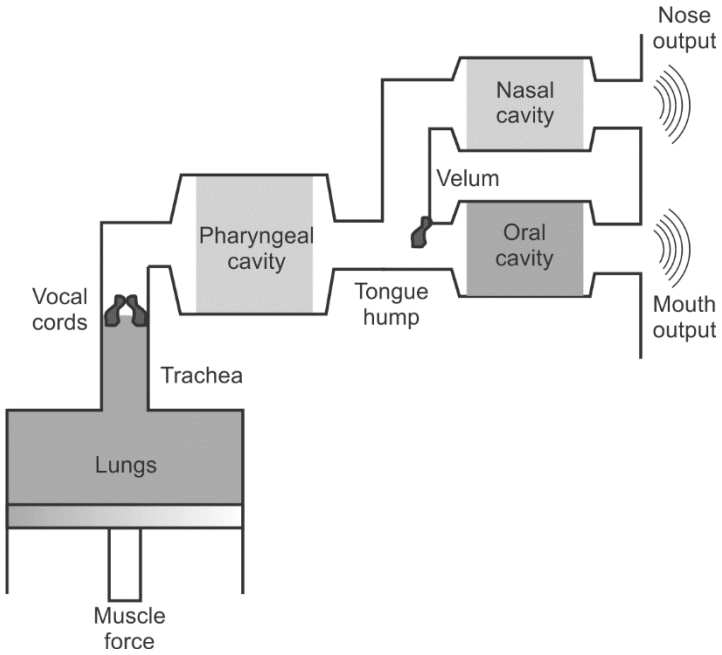


Fig. 1.2: Schematic representation of complete physiological mechanism of speech production.

Source: Anwary, (2004).

The representation of Speech by three states is:

- Unvoiced (U) when the vocal cords do not vibrate; hence the resultant waveform is periodic or random in nature
- Silence (S) when no speech is generated
- Voiced (V) in which the cords of the vocal cords are tensioned and consequently vibrate regularly if the air flows from the lungs.

Digitizing Speech

The initial phase in the speech recognition process is to transform analog representations into a digital signal (the first air pressure, then an electric analog signal in a microphone). Fig. 1.3 depicts the analog to digital speech conversion stages. There are two stages in the process of analog to digital conversion: (Digitization). The sampling frequency is the

proportion of samples obtained in the second by measuring their amplitude at a particular time. A wave must have at least two data in each cycle for reliable measurement: one measuring the wave's positive half and one measures the adverse component [4].

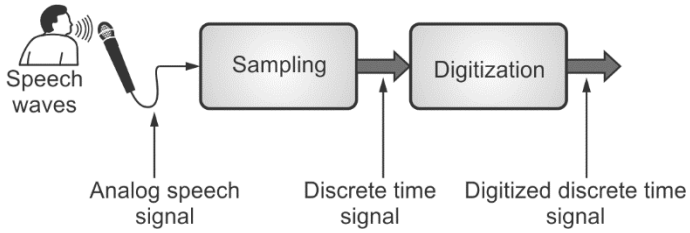


Fig. 1.3: Digitization of Speech

Source: ALTHOS, 2009

More than specimens improve each cycle's amplitude precision, yet the wave's speed is increased by less than two. Therefore, one with the frequency of half the sampling rate is the most significant wave which can be observed (since every cycle needs two samples). This maximum frequency is known as the Nyquist frequency for a specific sampling rate. Most human-speaking information is less than 10,000 Hz in frequencies; a sample rate of 20,00 Hz is thus needed for complete accuracy. The sense amplifier nonetheless filters telecommunications. The frequency of telephones is less than 4,000 Hz to be broadcasted. For telephone bandwidth speech, like the Switchboard corpus, the 8,000 Hz sample rate is adequate. The 16,000 Hz (also dubbed broadband) sample rate is widely utilized for microphone speech [6].

A predictive performance model must be constructed based on numerous criteria found to generalize the employment of such systems in diverse man-machine interaction scenarios. As voice recognition research is underway every day, researchers concentrate on improving performance and developing business applications in real-time. The fundamental objective of Speech Recognition research is now to enable a computer to detect words in real-time, with 100% accuracy.

This Speech technology automatically detects gaps between the words and divides the voice stream into independent words without explicit user signals. An end pointer to detect speech/silence limitations is termed a speech/silence detector (shown by dotted lines). With

the choice of speech segments between the intervals of succeeding quiet segments, words may be separated [7].

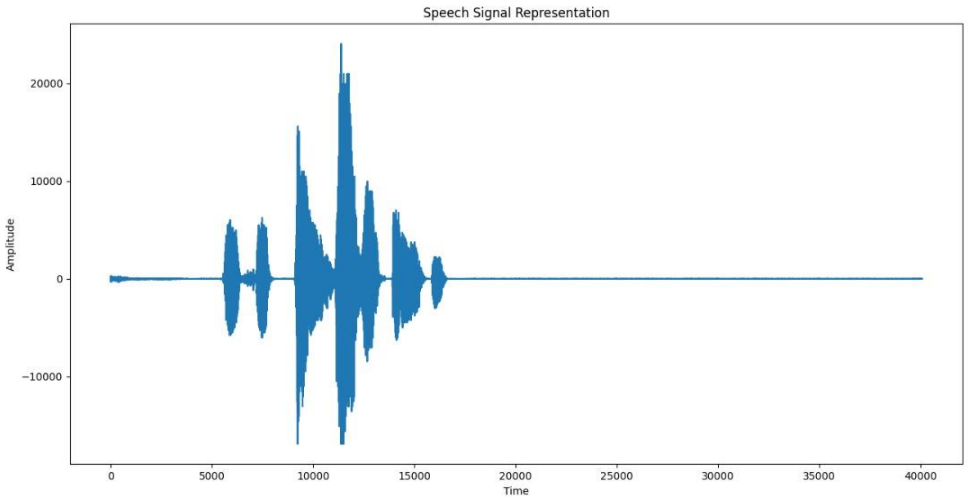


*Fig. 1.4: Representation of speech signal "*कृषी उत्पन्न बाजार समिती*"*

A basic end-point strategy is built on sections of quiet, which have low signal amplitude. For example, the representation of the voice signal " कृषी उत्पन्न बाजार समिती " is shown in Fig. 1.4.

1.1.1 Overview of Speech Recognition System

Speech is a natural language spoken by a human being for communication. The technical definition of Speech is an amplitude, frequency, and multi-component signal, with varying times. Because of this heterogeneity, transitions in distinct frequency bands may occur at various times. A speech recognition system is a technique used to extract, recognize and translate the speaking features utilizing intelligent electronic equipment. The System's primary objective is to build a technology for human interaction, independent of vocabulary, sound, Speech, or accents, in our natural language in real-time. For single-word recognition, continuous speech recognition, and spontaneous speech recognition, this System may be applied. It has used for young children, for telecommunications, for persons with hearing impairment [8].

Speech recognition systems like Google Voice Search that support around 120 languages, made remarkable improvements in recent years. It is important to academics and businesses to extend its coverage to global languages[8]. Speech Recognition systems for Voice Detection, Communication, and Control systems have increased enormously to improve the user experience and make the System effective and efficient. The critical phases involved in speech recognition systems are pre-processing feature extraction and classification/ pattern recognition. Due to the surrounding noise and other kinds of problems, disorders in speech signals increase, affecting the performance of the speech recognition system. Accuracy of recognition is utilized in the speech recognition system for the performance assessment [8].

The microscopic study has been carried out in the case of Indian languages compared to non-Indian languages. As extensive research has been done in other developed nations about Speech Recognition Systems, the quantity of work done in Indian regional languages has not yet reached a threshold level that makes it a meaningful communication tool. Marathi is a language spoken in western and central India, i.e., Maharashtra and some parts of Madhya Pradesh, Gujrat, and Karnataka. Ongoing research is, in a way, on the isolated identification of words spoken in local languages in various locations of India. Speech is the most noticeable and natural mode of interpersonal communication. The globe is conceived of in many spoken languages. There are, nevertheless, many possibilities for developing systems that use Indian languages that vary[9].

Words recorded by the speaker, the microphone, and the telephone converted to an auditory signal are the primary purpose of language recognition. An area of the ongoing investigation is isolated, extracting Marathi words to identify and validate each word uttered[9]. These Indian-Language systems are in their early years due to various obstacles, including resource deficiencies, despite the continuing improvement of Automatic Space Recognition (ASR) technology.

Some multi-lingual Acoustic Models (AM) need a standard telephone set. others include characteristics of the input noise. Speech recognition is the most advanced method. It employs neural network approaches based on voice recognition solutions. There has

already been significant development in this field; however, the System's resilience is essential. For real-time responsiveness, many training techniques are used. Most works on the identification of multi-lingual Speech were restricted to the multi-legalization of the acoustic model (AM) [10]. In [10]. the lower layers are shared throughout languages, with languages particular to the lower neural network (DNN) output layer. Alternatively, multi-lingual bottlenecks may be employed for either a Gaussian or DNN mixing model using a DNN function extractor. Since then, GMM and HMM have been used to construct many speech recognition applications. Later on, ASR was integrated into the machine. In addition to this, several Indian languages have been introduced, e.g., Bengali, Malayalam, Marathi, Hindi, Gujarati, Telugu, Bodo, Kannada, Punjabi, and Tamil. Researchers have recently developed a new methodology to assist more profound learning algorithms to simulate the spectrum fluctuations. Then, the use of profound learning in language recognition has grown enormously. Several techniques to deep learning have been documented, including deep belief (Baker, 2013), deep recurring neural networks [10], deep convolution neural convolution network (DCNN) (17)(18), and HMM hybrid Convolution Neural Network (CNN). However, ASR has a great deal to develop; Researchers are working to build an effective voice recognition system.

The objective of speech recognition is to create an ideal system for recognizing a sequence of words subject to language restrictions. The word is made of vowels and consonant linguistic units. A sentence model is supposed to be a succession of smaller unit models in recognition of Speech. The acoustic proofs of these components are paired with the principles for building valid, intelligible phrases.

1.1.2 History of Speech Recognition System

Research into the speech recognition system, including Microsoft Translators, Google Voice Search, Apple's Siri, etc., has substantially improved for 10 decades. This technology is currently employed in industry as well as in academics with tremendous interest[11]. Speech recognition is actively taken into account and has progressed enormously via digital signal processing in the voice recognition and communication

system. The Speech in different surroundings and noise during Speech greatly influences the Speech recognition system performance and accuracy.

Human beings employ a voice as an excellent communication channel. Therefore, we may describe Speech in simple words "Speech Recognition is a method that uses computer devices to extract, detect and interpret speech features." The development of these systems is intended to facilitate humans to operate products with a natural language, irrespective of vocabulary accent, language noise, and the surroundings. For the recognition of single, linked words, and continuous Speech, these systems are constructed.

In 1920 there was an acknowledgment of speech. The first machine was made, namely Radio Rex, a speech recognition toy. At the World Fair in New York, Bell Labs created a machine for voice synthesis. But subsequently, based on the wrong conclusion that AI is finally necessary for success, they abandoned attempts. In the 1950s, researchers explored phonetic acoustic ideas in their development of speech recognition. Most of the speech recognition systems in 1950 analyse the spectral resonances of the vowel in each syllable. A single-saver digital recognition system is anticipated by formant frequencies calculated in the vowel regions of each digit at Bell Labs, Davis, Biddulph, and Balashek (1952). Forgie & Forgie builder-independent 10-vowel recognition at the MIT Lincoln Lab in 1950 by detecting vowel spectral resonances. RCA Labs, Olson, and Belarus (1950) developed ten syllable recognitors with one vowel. In 1959, Fry and Denes (1959), by use of a spectrum analyser and pattern match, sought to develop a phoneme identifier to identify four vowels and nine consonants in University Collège in England. In the years 1960-70, Japanese laboratories entered the area of recognition. Computers are not fast enough, therefore they have developed H/W as part of their system for a particular purpose. A Radio Research Laboratory system, a H/W vowel recognitor, disclosed in the Tokyo, Nagata et. al. Sakai and Doshita from Kyoto University developed a HW Phoneme Recognizing Unit in 1962. Another attempt was made. Nagata and NEC Labs colleagues developed a digital recognizer in 1963 [1]. This resulted in an extensive and fruitful research programme. In 1970, isolated word recognition was the main focus of study. Wide recognition of spoken vocabulary was investigated by IBM researchers. Researchers started voice recognition

studies independently of speakers at AT&T Bell Labs. A wide variety of clustering methods have been used to determine the number of different patterns necessary for word recognition. This study is improved to make widespread use of the methods for speaker-independent models. The Harphy system of Carnegie Mellon University acknowledges speech with acceptable accuracy in the lexicon of 1011 words. It was the first to operate a finite state network to minimise calculation and efficiently identify matching strings. The main emphasis of the study in 1980 was on the identification of linked words. In the early 1980s, Moshey J. Lasry investigated letters and numbers in voice spectrograms and created a feature-based recognition of speech. In 1980, technological changes in particular for HMM in speech research from template-based methods to statistical modelling. Statistically, particularly stochastic processing with HMM [12] was introduced in the early 1970s, the most important paradigm change [13]. This approach still prevails more than 30 years later. Despite their ease of use, the language models of N-gram have proven remarkable. Most of the practical speech recognition systems today are based on a statistical method, and in the 1990s the findings have been further improved. One of the major technological developments was the Hidden Markov (HMM) method in 1980. The HMM was understood by IBM, IDA, and Dragon Systems, but not well-known in the mid-1980s. Another technique reintroduced in late 1980s is neural networking for voice recognition issues. 1990 saw the development of a pattern recognition method. Bayes typically followed the framework; however, this was modified to reduce the mistake in empirical recognition to an optimization issue. This change is because the spoken signal distribution functions cannot be precisely selected, and Bayes' theorem cannot be used under such circumstances. The goal, however, is to build a recognizer that is not the best matching data with the least recognition error. Minimum classification error (MCE) and Maximum Mutual Information are the methods used to minimize errors (MMI). These methods have led to a high probability approach to the execution of voice recognition. A weighted HMM method was suggested to solve problems of robustness and discrimination based on HMM voice recognition. A maximum probability stochastic matching method was suggested in order to reduce the acoustic malfunction of a given set of speech models

to test utterances. A HMM vocal recognition system storytelling method builds upon the usage of the neural network as a vector quantizer, which represents a significant breakthrough in neural network training. Nam Soo Kim et al. described many techniques for assessing a resilient HMM distribution of probability. The expansion of the Viterbi algorithm makes the second-order HMM as effective as the Viterbi algorithm already in existence. DARPA continued to run throughout the 1990s. The emphasis is then on the Air Travel Information Service (ATIS) job and subsequently on broadcast news transcription (BN). Progress has been explained in continuous speech recognition and noisy speech recognition. The little amount of effort was done on the loud speech. A novel method to an auditory model has been suggested for noisy settings to establish a robust speech recognition. Compared with other models, this method is computer-efficient. A spectral assessment method was created based on a model. A variable Bayesian method of estimation was proposed in 2000. It is based on the subsequent parameter distribution. Giuseppe Richardi has created an ASR adaptive learning solution method. In 2005, there were several enhancements to the mechanism for the identification of performance by large vocabulary on an ongoing basis. The Corpus of Spontaneous Japanese (CSJ), a five-year nationwide initiative was carried out in Japan. It contains about 7 million words, equivalent to 700 hours of speech. Acoustic modelling, identification of sentence limits, modelling of pronunciation, adaption of acoustic and language design and automated speech synopsis are methods utilised in this research. Utterance tests are under study to further improve the robustness of voice recognition systems, in particular for spontaneous speech. They utilize multimodal communication when people talk to one another. When communication takes place in a loud setting, this improves the rate of successful transmission of information. The use of visual facial information, particularly lip movement has been investigated in speech recognition and findings indicate that both methods provide higher performance than utilising audio or just visual information, in particular in the noisy environment[14].

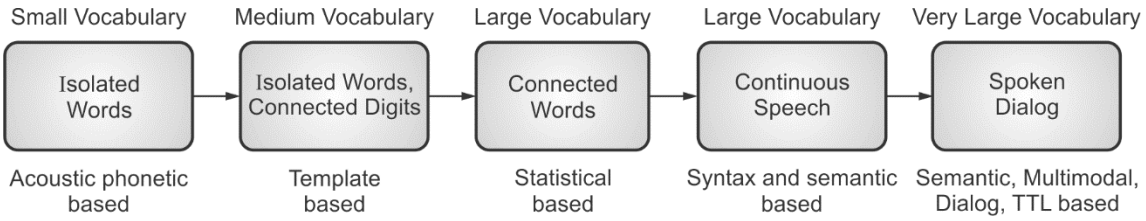


Fig. 1.5 History of Speech Recognition System based on vocabulary

Fig. 1.5 shows the journey of Speech Recognition system design. It started with a very tiny vocabulary in the identification of a single word. The purpose of this approach was to identify relatively few individual words. This concept was based on phonetic and acoustic vocabulary. H. Belar et al. developed a later Phonetic Typewriter in 1957. (2016). Later on, the template was created for the voice recognition system. These systems can detect single words and related numbers. These ASR systems employ the vocabulary of medium size. Then a statistically based ASR system and these huge vocabulary ASR systems were introduced. Because of this, linked words may be recognized by these systems. Later, the development of the ASR system is based on statistical models. The associated terms and vocabulary might be seen in these systems. Yet, the only answer for the ASR system is the acknowledgment of related words. That's why a large vocabulary has been created into a new ASR syntax and semantic-based System. Following that, an ASR system was created that supports the spoken dialogue. The System builds on semantical, multimodal, and TTS dialogues. In this ASR system, the system is implemented with a vast vocabulary [15].

The Speech Recognition System was widely accepted and used when the world's famous IT firms, including Facebook, Google, Amazon, Microsoft, and Apple, began giving this feature in different devices via services such as Google Home, Amazon Echo, Apple Siri, and many more. The objective of these leading technology businesses is to respond and answer voice helpers more accurately [15].

1.1.3 Types of Speech

- **Isolated Words**

Isolated reading comprehension is ideal for instances in which the user only has to offer one-word answers; however, it is pretty unnatural for many text inputs. It is easier to use

since word borders are apparent and the terms are usually fully defined, which is the main benefit of this kind.

- **Continuous Words**

A system of associated words is like an isolated one but allows for "run-together "independent utterances with a bit of interval between them. An utterance is the vocalization of a single word or word for the computer.

- **Continuous Speech**

Continuous Speech enables people to talk nearly usually while the computer controls their content. It's computer dictation, essentially. Without any break or any other divide between words, these closest words flow together. As a result, it is challenging to design an ongoing speech recognition system [16].

- **Spontaneous Speech**

The System of spontaneous speech recognition identifies the natural language. Spontaneous Speech, which abruptly enters the mouth, is natural. A spontaneous speech ASR system can handle a range of natural speech properties such as words. Spontaneous Speech may involve mispronunciation, mis beginning, and nonspeaking.

1.2 Types of Speech Recognition Systems

Due to their distinctive physical form and personality, every speaker has a distinct voice. Therefore, a system of speech recognition is characterized as follows in three major categories:

Speaker Dependent System

For a certain kind of situation, speaker-dependent systems are designed. They are typically more precise for the individual speaker but could be less precise for other speakers. As a result, these systems typically are less expensive, simpler to design, and more precise. But as speaker-independent systems, these technologies are not adaptable.

Speaker Independent System

Speaker Independent System can detect different speakers without prior training and previous instructions. For every specific kind of speaker, a speaker-independent system is

designed. It is used in the IVRS system, which requires the access of a wide range of users to get information. However, the disadvantage is that it restricts vocabulary terms. The Speaker Independent system is the hardest to implement. It is also pricey and less than speaker-dependent systems its precision [9].

1.3 Understanding Marathi Language

The Marathi language belongs to the family of Indo-Aryan language, which is a part of the biggest group of Indo-European languages. Marathi is the official language of Maharashtra state in India. The Sanskrit language originates in all Indo-Aryan languages. Three languages from Prakrit, i.e., the Saurseni, the Magadhi, and Maharashtrian languages, come from the Sanskrit language. The Devanagari script is used in the Marathi language. There are around 72 million Marathi speakers worldwide[17]. The Marathi language has a total of 12 vowels and 36 consonants. “There is a specific Marathi language department at institutions like Maharaja Sayajirao University, Baroda, University of Osmania, Andhra Pradesh, Gulbarga University, Devi Ahilya University, Indore Universities, and Goa University, Panaji for studying a Marathi Language and its literature”. “Hindi is written in Devanagari while the other 17 Indian constitutionally recognizable languages are (1) Assamese (2) Tamil (3) Malayalam (4) Gujarati (5) Telugu (6) Telugu (7) Urdu (8) Bengali (9) Sanskrit (10) Kashmiri (11) Marathi is thought to be the descendent of Maharashtri, spoken by Prakrit in Maharashtra area”. In addition to Sanskrit, Marathi was also impacted by Kannada (state of Karnataka) and Telugu languages in its bordering states (Andhra Pradesh). The actual Marathi script is termed 'Balbodh,' an altered Devanagari form. A script named 'Modi' was used before Peshwas's reign (18th century). Hemadpanta, a minister in the courts of the kings of Devgiri, introduced the script (13th century). This alphabet looked more like the current Dravidian scripts and benefited a higher written speed, as the letters could be combined. Today, only the Sinhala script is utilized that is simpler to read and has no benefit in writing quicker.

The character set of Marathi

Marathi is mainly written in a script that's phonetic in nature called Nagari or Devanagari. The characters are classified into Vocals and consonants. Every vowel has its symbol in Marathi. In the Marathi language, there are 12 vowels. The consonant itself has a vowel + (too) implicitly. A vowel sign (Matra) is appended to the consonant to indicate a non-implicit vowel-based sound. Fig. 1.6 shows the list of Marathi consonants

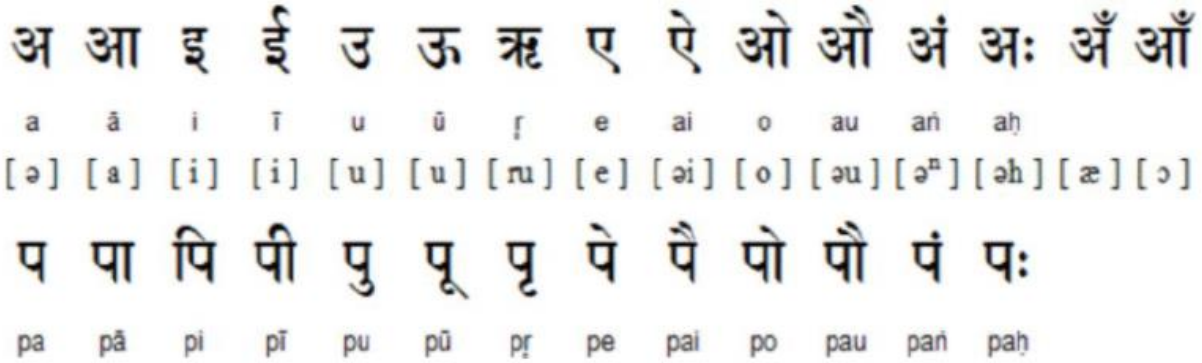


Fig. 1.6: Vowel Set of Marathi

Consonants: The consonant in Marathi is classified by location and way of articulation in several categories. The 5 Vargs (groups) and 9 non-Varg consonants are split. There are five consonants in each Varg, the last being a nasal one. The primary and secondary pairs of the first four consonants are each Varg. The primary consonants are voiceless, whereas minor consonants are voiced. Each pair's second consonant is the aspired counterpart (the sound is an "h" extra). There are so four consonants of each Varg (unvoiced), (unvoiced, aspired), (voiced), respectively (voiced, aspired). 9 other non-varg consonants are grouped into five semivowels, three wheelchairs, and one aspirate. The whole phonetic property set of Marathi consonants is shown in the Fig. 1.7.

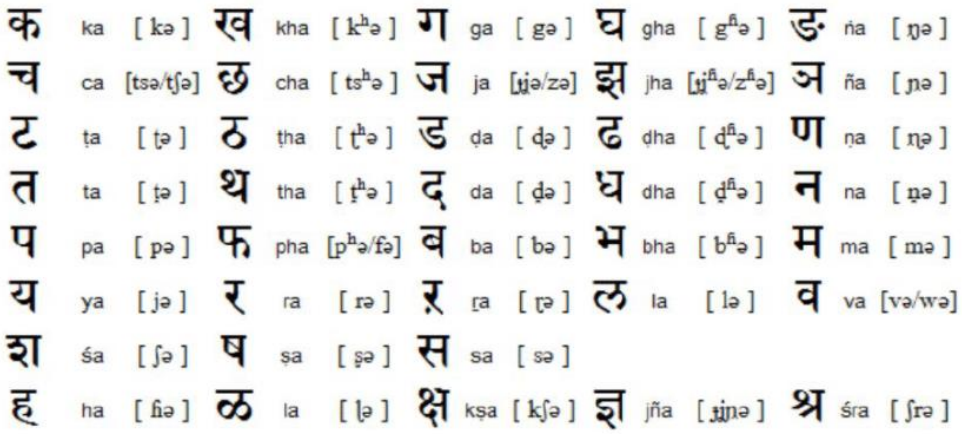


Fig. 1.7: Marathi Consonant Set

Other Characters: "Apart from consonants and vowels, there are some other characters used in the Marathi language are anuswar (◌ं), visarga (◌ः), chanderbindu (◌ँ), >, s, @, ◌ॉ. Anuswar indicates the nasal consonant sounds. Anuswar sound depends upon the following character. It sounds wisely depicts the nasal vowels of these vargs depending on the varg of the following letter".

1.4 Research Motivation

Several researchers, such as Marathi, Punjabi, Kannada, etc., have lately started investigating a voice recognition technique[18]. Many research has demonstrated that background noise is a significant problem when a Speech recognition system is designed. People in challenging acoustic environments are skilled at identifying Speech. ASR systems are now built in a very excellent environment or utilizing a microphone[19], however, ASR systems are not dependable and competent in the field of human voice recognition when Speech signals noise increases.

The world is moving towards inclusive development in all aspects of geography, culture, and many more. When we say inclusive design, we can not force anyone while using the product to use this language or that. It is an approach to design a product in a such way that whoever wants to use the system or product can use it the way they want. The best example of it is our mobile phone. Now, the mobile phone can be accessed in almost all regional

languages which allows uneducated people to have access to the technology. The voice interface is now an emerging user interface. We must consider all the regional languages and its dialect to make the voice interface more inclusive and accessible technology to all the people irrespective of their educational backgrounds, culture, and language, etc. and this is the research motivation behind this thesis.

1.5 Contribution of Thesis

We have conducted a detailed technology study on Speech Recognition systems and research in Indian languages to date. First, we contrasted the technology of machine learning with the neural network. Then, using the Hybrid RCBO technique, we optimized the Deep Belief Network. RCBO-DBN is then compared to CNN, RNN, LSTM, followed by a ROA-CNN comparison. Finally, an LSTM approach based on Big Bang Big Crunch has been tested.

1.6 Organization of the Thesis

- Chapter 1 Introduction discusses the Types of Speech Recognition Systems, Research Motivation, and Contribution of the Thesis.
- Chapter 2 literature surveys, different approaches used for developing speech recognition systems are discussed. After that, research in developing a speech recognition system for Indian Languages is discussed. Then Research Gap, Research Problem, Research Objectives are discussed.
- Chapter 3 Research Methodology explains the Machine Learning Techniques, Deep Learning, and Speech Dataset.
- Chapter 4 Deep Belief Network using Duo Features with Hybrid-Meta-Heuristic approach for developing Marathi Speech Recognition System
- Chapter 5 is Enhanced Marathi Speech Recognition enabled by Grasshopper Optimization-based Recurrent Neural Network
- Chapter 6 is Big Bang Big Crunch Based LSTM Approach for Developing Marathi Speech Recognition System

- Chapter 7 Compares proposed approaches during the research
- Chapter 8 Conclusion and Future Scope explains the Conclusion and Future Scope of the study.

1.7 Summary

This chapter has presented an introduction to studying the design and development of a speech recognition system for the Marathi language. Moreover, the chapter introduces an Overview of the Speech Recognition System, the History of the Speech Recognition System, Speech Recognition System Architecture, Types of Speech, Types of Speech Recognition Systems, Research Motivation, Speech Recognition Process, Contribution of Thesis, and Organization of the thesis. The next chapter will present the Literature Review of the study.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Literature review summaries the overall available information and findings on the problem/ domain/ topic we are working on. It helps to analyze the past work done on the genre and its limitations to the present. The review indicates further research that needs to be done on the topic. The earlier writings and information provided are analyzed, and the authority that has done good work should be acknowledged. Such type of reviews helps the researchers to understand the background to the further research that needs to be done on a particular topic. Such discussions must have a summary and must highlight the areas of disagreement and agreement, and the pointing out of the gaps must be done.

A proper literature review is always required so that the overall knowledge or information on a topic can be gathered and determine the areas that have been well highlighted and the areas that need further highlighting.

While performing a literature survey, we have considered the below-mentioned points –

- Need of Speech Recognition System (SRS) and its development history
- What are the different techniques used for developing Speech Recognition System?
- Research work done in developing SRS systems for Indian Languages

2.2 Approaches for Developing Speech Recognition System

The design and development of Speech Recognition Systems have been implemented using different approaches. There are three main approaches discussed here to understand the relevance of these approaches for developing the Marathi Speech Recognition System shown in Fig. 2.1. It includes the Acoustic-Phonetic approach, Pattern Recognition Approach and Hybrid Approaches.

In recent years, pattern recognition was solved using dynamic programming approaches[20]. The Hidden Markov Modeling (HMM) technique was integrated with stochastic modeling methods to handle speech/voice recognition. Further study was based

on the Artificial Neural Network (ANN) concepts, in which biological neural systems simulate the computing of features extracted. Much current research on speech identification involves continuous vocabulary recognition utilizing HMMs, ANNs, or a hybrid form[21].

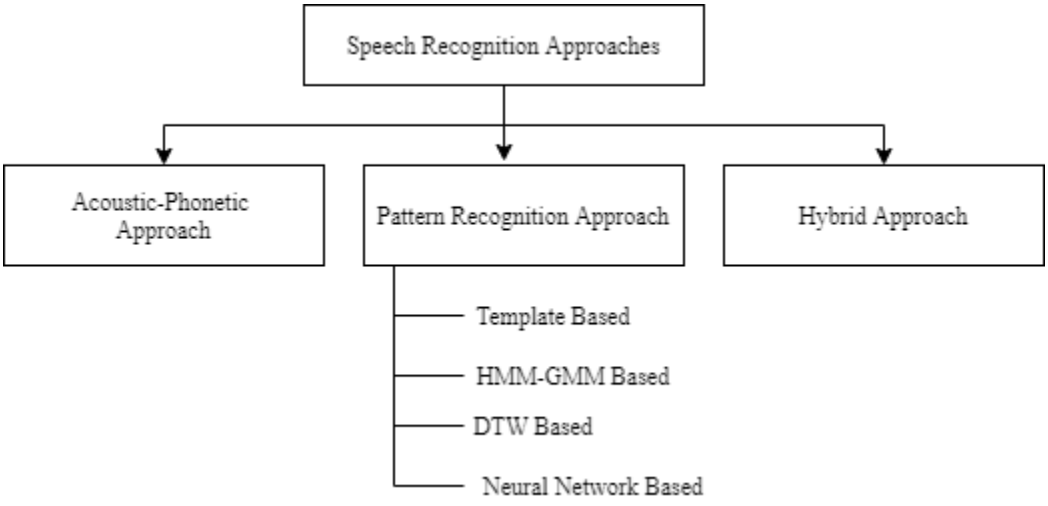


Fig. 2.1: Speech Recognition Approaches

2.2.1 Acoustic Phonetic Approach

The acoustic phonetics suggests that in spoken language, there are limited, discrete phonetic units. The phonetic units are characterized by several acoustic characteristics seen throughout time in or within the speech signal. The spectrum analysis of Speech and features detection, which transforms spectral data to a collection of characteristically broad-based acoustic characteristics of the various phones, is the initial stage in the acoustic phonetics technique. The following stage is a segmentation and labeling phase during which the voice signal is split into stable acoustic areas, and one or more phonetic labels are attached to each segmented area, which results in a characterization of the phoneme lattice of the Speech. The last stage in this methodology is to identify a legitimate word (or word string) from sequences created by the phonetic label segmentation to labeling [22].

- **Acoustic Model**

The acoustic model is an essential source of information for speech recognition systems, acoustic properties for phonetic units. In creating an acoustic model, the selection of basic modeling units is a fundamental and significant topic. In general, several sorts of units can be utilized to represent it acoustically when specifying the Speech. The efficacy of speech processing may be significantly affected by different acoustic modeling units.

Acoustic speech modeling often refers to a technique of statistics for feature-vector sequences calculated from the speech signal waveform. One of the most prominent statistical models to generate an acoustic model is the Hidden Markov Model (HMM). Segment models, super-sectional (with hidden simulated data), neural networks, maximal entropy model, conditional random fields, etc., are also part of other acoustic models. An image representation is a file containing statistical estimates of each sound that constitutes a word. A phoneme mark is given for both statistical representations. A vast speech database termed a speech corpus and unique training techniques to build statistical representations for each phoneme in a language are used to develop the acoustic model. There is an HMM in every phoneme. The speaker listens to the different sounds that a user speaks and then searches for an appropriate HMM in the acoustic model. Every word being uttered is broken down into an order of simple sounds known as fundamental phones. The acoustic model explains the likelihood of a particular observation with a base phone.

- **Language Model**

A language model is a set of restrictions on the sequence in a specific language of allowed words. For example, the rules for generative grammar or just the statistics on every word pair predicted on a training body might reflect these limits. Although terms have the same sound telephone, it is often not difficult for people to distinguish the term. They know the context and are also relatively aware of what words or phrases might arise. The objective of the language model is to provide this context for a voice recognition system. The language model defines which words are valid in the language and which order.

Language models normally can calculate the n-gram probabilities by watching corporal sequences of words that generally comprise millions of word touches and eliminate

confusion in data formation. However, it is noted that less confusion does not inevitably equate to improved outcomes in spoken recognition. There are thus desirable algorithms that enhance language models depending on their influence on speech recognition. A language model that describes the distribution of likelihood of words, given the history of said word, that the speaker may pronounce next. Bigram and trigram are standard language models. These models provide calculated probabilities in a succession of groups of two or three specific words. Language Modelling Toolkit (SLM) from Stanford Research Institute Language Modeling Toolkit is available.

- **Decoder**

Due to the input pattern 'I' and the acoustic-phonetically speaking model, the decoding step aims to discover the most probable word sequence W. Dynamic programming approaches help tackle the decoding challenge. The emphasis is on identifying a single route across the network which is the best match for 'I,' instead of assessing probabilities of all alternative model routes producing I. The Viterbi method is often used to determine the optimum state sequence for the observation series [21]. For more extensive vocabularies recognition problems, the Viterbi algorithm's recurring phase might be problematic to evaluate all possible words. To remedy this, a beam search for a Viterbi iteration can only be employed if you extend the routes to your next step, words with route probabilities over a threshold. This technique accelerates the search process at the price of decoding precision. The Viterbi method expects each of the best pathways at 't' should be an extension of the best pathways that finish at 't – 1', which is typically not true. In the first place, the routes which look less likely than others might become the ideal route for the whole series (e.g., the most probable phoneme sequence does not need to correspond to the most probable word sequence). Extended Viterbi and forward-looking algorithms resolve this problem.

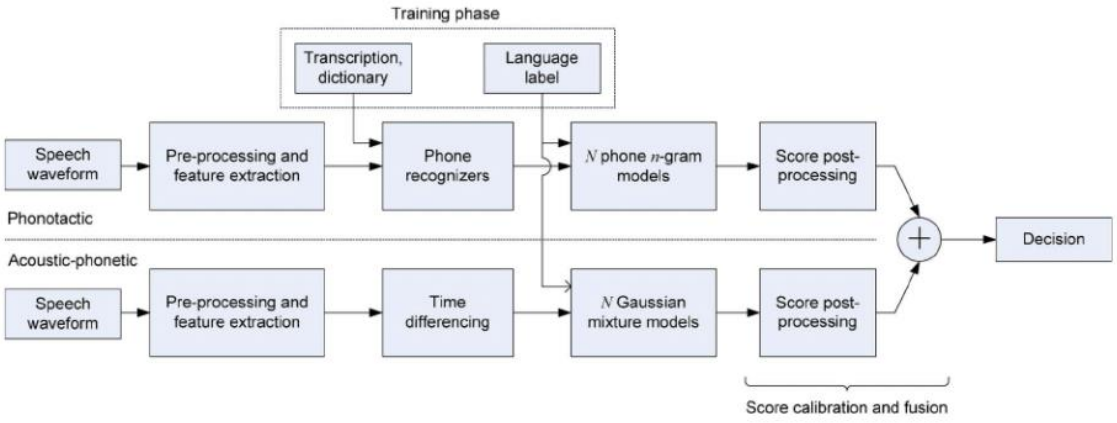


Fig. 2.2: General scheme of acoustic-phonetic approach

2.2.2 Pattern Recognition Approach

In general, Speech Recognition Systems have 3 steps, processing, extraction of features, and decoding in this approach. Again, decoding the sound model, the language model, and the pronunciation model carried out their work.

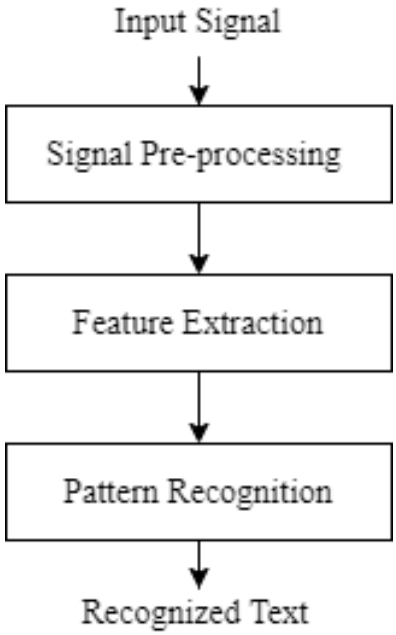


Fig. 2.3: Block diagram of pattern recognition approach

Speech is a sound sequence with several related qualities. First, the noises are transformed into tiny frame fragments. The extraction of features analyses these frames to derive speech vector features. Language recognition is similar to the System of pattern recognition.

Pre-processing: A digital device cannot directly handle speech signals because they have an analogue waveform. As a result, the incoming speech is pre-processed before recognition. To begin, voice recognition is digitized for this purpose. First-order filters flatten the spectral response of the digital voice signals, which have been sampled. Pre-emphasis is a technique that reduces the size of higher frequencies to the magnitude of their lower counterparts. As a further step, you must block the voice signal by using an overlap range of 50 to 70 percent of frames to block the signal. The frame size range should be somewhere between 10 and 25 milliseconds.

Feature Extraction: The objective of extracting features is to detect a collection of characteristics of a word that have acoustic correlations with the spoken signal, parameters that may be calculated or inferred by processing a signal waveform. These parameters are called functionality. The extraction procedure should eliminate unnecessary information while retaining important information. It comprises the measurement of crucial signal features, such as energy or frequency response (i.e., signal measurement).

The primary objective of the feature extraction process in recognition of the Speech is to calculate a parsimonious series of feature vectors to describe the input signal presented compactly. The extraction feature usually takes place in three phases. The first step is termed an acoustic front or a speech analysis. The spectrum analyses the signals temporarily and creates raw characteristics that describe brief spoken intervals' energy spectrum envelope. In the second step, an expanded vector with static and dynamic characteristics is compiled. In the penultimate step (which does not always exist), these extended functional vectors are transformed into compact and resilient vectors that the recognizer subsequently provides.

There is nothing suitable for particular objectives. However, the selection of features has the following characteristics: these may allow an automated message to discriminate between sounds with similar sounds, they should allow the automatic creation, without the

need of an excessive quantity of training data, of feature vectors for these sounds and it is vital to use processes to reduce the information in every segment of the audio signal to comparatively a limited number of characteristics or feature to locate certain statistically valid information from incoming data. These characteristics should characterize each segment so that additional comparable segments may be merged by comparing their characteristics. The speech signal in parameters is enormously intriguing and remarkable. Some of the extraction methods for the feature include Linear Discriminant Analysis (LDA), Linear Predictive Coding (LPC), Principles Component Analysis (PCA), Cepstral, Dynamic feature extraction, wavelet-based features, Mel-Frequency Cepstral Coefficient (MFCC) etc.

In the field of noise-robust speech recognition, several auditory feature extraction processes are used, for example, zero-crossing peak amplitude, mean-localized synchronous detection (ALSD), Perceptual Minimum Variety Distortion-less Reaction (PMVDR), Power Normed Cepstral Coefficients (PNCC).

Pattern Recognition

1. Template Based Approach

The template-based method features a prototype pattern collection. These patterns are saved as reference patterns for the word dictionary. Speech is detected by combining an unknown oral expression with these reference templates and picking the best-matched pattern category. Usually, Complete word templates are built. However, the errors may be prevented by segmenting or classifying smaller acoustically more changeable components such as phonemes [23].

Template-based speech recognition methodology has offered a family of technology that has made considerable progress in the previous two decades. This is a straightforward technique. It coincides with unknown speeches versus pre-registered terms or templates to find the best match[24]. This strategy benefits from employing exact word models but has the disadvantage of fixing the pre-registered templates. Variations in the voice signals may thus be represented only by the use of numerous templates per syllable. Training and

matching templates are prohibitively costly or unworkable when vocabulary is more than a few hundred words[25]**.** In terms of storage and processing power required to accomplish matching, this strategy is relatively ineffective. Matching templates is also reliant on tiresome speakers. This technique cannot be used to recognize continuous speech.

2. Neural Network-Based Approach

Deep learning is a new field of machine learning, sometimes known as representational learning or unattended feature learning. Deep learning is becoming a popular speech recognition method that has successfully substituting Gaussian language recognition and feature coding on an ever-greater scale [26]. The first category comprises deep generative architectures designed to characterize the high-quality correlation qualities of observable data and their associated classes, or joint statistic distributes. However, this sort of design may become discriminatory with the use of the Bayes rule. Examples include numerous deep auto encoding forms, Boltzmann machine depth, sum-generated networks, the original shape of the Deep Belief Network (DBN), and its extension to the Boltzmann machine in its base layer[27].

3. HMM-GMM Approach

The SRS is based on the most common generative learning method, Hidden Markov models. Conventional voice recognition systems utilize the Hidden Markov model-based sequential structure of GMM-HMM[28]. The use of HMMs is partial as a stationary signal or short-time stationary signal for language recognition. Speech may be modeled in a short period on a stationary procedure as a Markov model for many stochastic reasons[9][29]. In typical cases, each HMM state uses a Gaussian mixture to shape the sound wave spectrally[30]. AGMM-HMM is parameterized by the probabilities (A, B, μ), [31] which are the Gaussian mixture model of state (J). AGMM-HMM is the state transition probability matrix (A, B, μl). The condition is usually related to a spoken telephone subsection[32]. The State-of-the-art systems are based on performance levels. The HMMs are popular because they can easily manage sequences of varied length data resulting from changes in word sequence, speech pace, and accent. However, the HMM-GMM technique had become the standard tool for the drawbacks of ASR it has its own. HMMs-based

language recognition systems may be taught automated and easy to use. One of the critical disadvantages of Gaussian blending models though is that they are statistically ineffective for modeling data on or close to a non-linear manifold in the data area.

4. DTW Approach

Dynamic Time Warping technique (DTW) results in the optimum warping of 2-time series. Dynamic Programming (DP) in DTW is used to identify the lowest distance trajectory to lower the quantity of the computation. Time sync is utilized for dynamic programming where every time matrix column is taken into account.

There were two primary ways to matching patterns in ADA: deterministic matching patterns based on Dynamic Time Warping (DTW) and stochastic matching of patterns based on the usage of Hidden-Markov Models (HMMs). In DTW, one or more templates represent each class to be identified. Then order to improve the pronunciations/speaker, variability modeling might be more desirable than one reference template per class. A distance from the observed sequence of Speech and class patterns is determined during recognition. The detected word matches the route across the models, which minimizes the distance accumulated. The stretched and boogied versions of the reference patterns are also used in the distance computation to reduce the duration malfunction between the test and reference patterns. Increased class patterns and removal of warping restrictions might enhance the DTW-based performance in storage and computer needs. Compared to the DTW duet, improved generalization and memory needs are favored with HMM-based pattern matching[30].

2.2.3 Hybrid Approach

The usage of neural networks is another way for matching patterns in the voice reconnaissance system. Neural networks can do sophisticated recognition tasks but can't perform well if there is a vast vocabulary [33]. They can handle poor quality, noisy data, and the independence of speakers. When training data are available, and vocabulary size is constrained, this sort of System may attain more precision than HMM systems. Phoneme identification is a well-known technique employing neural networks. This is a dynamic

field, yet its results are often better than those of HMMs. Also, an NN-HMM hybrid system is available that employs the neural network to recognize phonemes and HMM to represent the language.

A. The Modeling Unit generally decreases the recognition rate for the whole System by increasing phonemes' recognition accuracy in the Phoneme Modelling process.
B. Different features are taken of the speech signal, the hybrid network model (HMM+NN), and a range of knowledge sources, including characteristics, vocabulary, and meaning, to comprehend research to improve systems attributes to promote speech recognition.

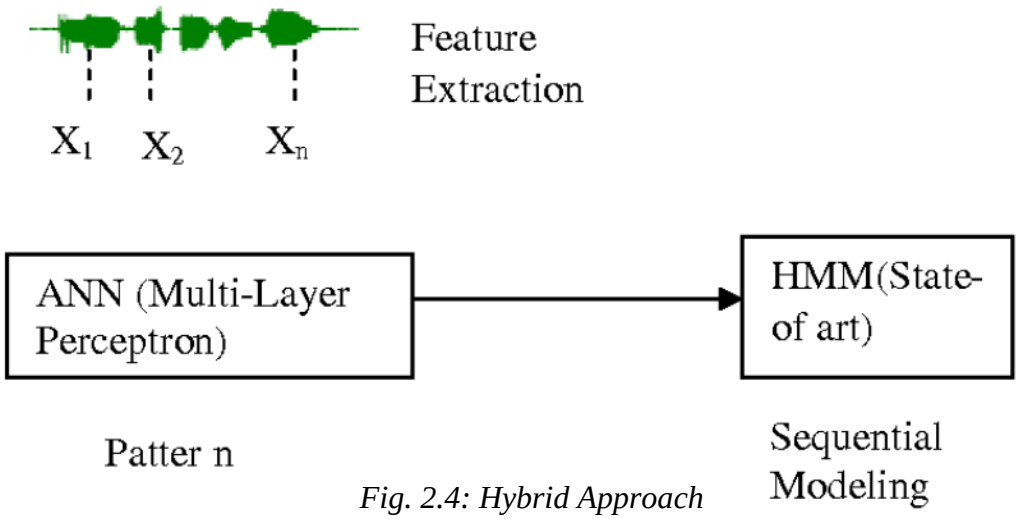


Fig. 2.4: Hybrid Approach

In recent years there has been considerable growth in using the deep neural network in language recognition. The following areas may be separated into convolutional neural networks in the speech recognition process: First, the efficiency of neural nets is improved. Second, a combination of a hybrid system may be developed. Thirdly, the unique character of the neural network and mathematical approaches is applicable in speech recognition. A new growing trend has become an artificial neural network in voice recognition. Applied effectively to solving pattern classification issues [34] with the use of artificial neural grid technology, it has shown enormous energy, speech recognition systems that use the artificial neural network are on the market, and people are going to change their approach to suit a wide range of identification systems.

Approach to knowledge

Several scholars have advocated using the knowledge/regulatory method for language recognition and applied Speech recognition. A technique used for knowledge-based methods are linguistic, phonetic, and spectrogram information[25]. The speaker takes several characteristics and then trains the System to develop a set of sample production rules automatically. These rules are based on criteria that give important categorization information. The effort of recognition is made to build the decision tree on the level of the frame, employing an inference engine to categorize the firing. This technique has the advantage of explicitly modeling speech variations, but this expert knowledge is hard to get and apply properly; therefore, this technique is seen instead as an impractical and automated learning process[35].

2.3 Review of Literature

In this section, the literary works based on speech recognition were illustrated along with the challenges and research questions.

Sun et al. [36] developed a multi-task learning framework using Deep Domain Adaptation (DDA) approach for speech recognition. Here, the label predictor was utilized to predict the phoneme labels and train the classifier. Then, the network optimization was carried out to reduce the loss of the label classifier and increase the loss of the domain classifier. The method was simpler to implement and improved the accuracy of the framework.

A new HMM-KNN hybrid classifier was used to develop a speech recognition system[37]**.** In this approach, HMM is used within KNN and trained in two different post-processing methods. The impact of model size parameters, number of layers, and architecture are associated with networks and training information such as loss function. Here, regularization approaches have been used to determine DNN classifier performance and to analyses speech recognizer word error rates. The approach uses a basic DNN framework and optimization methods for robust speech recognition[38].

A technique for large Hindi vocabulary is developed using the Kaldi automated speaker identification toolkit, called Context-Dependent Deep Neural network HMMs (CD-DNN-

HMM). By tweaking the DNN layer, the approach gives robust results for improvement. However, the approach was not relevant to the precise creation of DNN models using Hindi voice, and the auto-encoder and DBN could not be used to provide more precision.

Mao and Zhang et al. [39] developed a language-based recurrent neural network time-delay recognition system. This network comprises a neural network layer and recurring memory neural network layer. The entire network was then trained in a source language, and the hidden layers were then retrained using a tiny target language to enhance target language recognition performance.

Mitra et al. [40] devised a hybrid model named Hybrid Convolutional Neural Network (HCNN) for speech recognition. This model utilized two parallel layers for modeling the acoustic and articulatory spaces. Here, the articulatory information and filter bank-energy-level features were used for speech recognition. In addition, this model was utilized for performing time-frequency convolution to extract the time convolution on the articulatory features.

Lee et al. [41] designed a DNN-based ensemble structure for performing speech recognition tasks. Here, each DNN acoustic model was trained for mapping features and acoustic modeling. The ensemble of the DNNs is integrated with the weighted averaging of the prediction probabilities for identifying the noisy signals.

For language recognition utilizing data from big multi-speech vocabulary, *Kamper et al.* [42] have created a Segmental Bayesian Model. The method uses a Bayesian modeling structure employing segmentary word representations to divide each word by mapping the sequences of functional frames to one integrated vector, utilizing a fixed-dimensional acoustic embedding.

Xue et al. [43] have proposed a technique called the hybrid NN/HMM speaker adaption technique based on the uniquely decomposed value (SVD). This approach utilized SVD to the training of DNNs utilizing weight matrices. This approach has been utilized to alter the rectangular diagonal matrices utilizing the data of adaption. By updating weight matrices with the singular value, this technique addresses the over-service issue.

Murase [44] explored many methods, State Tracking Dialogue, knowledge base, graphic knowledge, and associative knowledge. The feature vectors are constructed in terms of association knowledge based on the knowledge graph. Tests in this research show the suggested efficiency of systems using the neural network. Based on the data, it is suggested that the optimal approach for DST is the functionality vector.

The RNN, CNN approaches, FACVS (knowledge vectors association), and FCNN algorithms utilized here have fully connected neural networks. It may be utilized to grasp the linguistic features of the inference approach offered to develop an associative feature.

A brief speech verification system based on profound neural networks has been suggested by *Guo et al.* [45]. In this study, the utterances of both males and females speakers are assessed using DNN1 and DNN2; as opposed to the methodologies of the papers, the suggested System has substantial performance. The performance of the GMM-DNN-based i-vector speaker grows with a length of utterance assessment.

The suggested deep learning approaches, presented by *Ming Li et al. (2011)* [46]**,** produce more outstanding performance in the segment, but not at the individual level. Deep learning helps to achieve optimum performance in the sector. Cross-validation technology improves accuracy at the level of individuals and sectors. The suggested approach provides better results not at the individual level but at the segment level to identify patterns in Speech.

A cross-lingual audio model based on temporal connectionist classification was introduced by *Sibo Tong et al.* [47] (CTC). When the phonemes are overlapped, this approach works efficiently. It works the same as it does for overlapping phonemes for new phonemes. The results presented in this research demonstrate that the CTC-based technique offered delivers substantial performance compared to the restricted data set DNN/HMM methodology. In this work, the acoustic, multi-lingual, and hidden unit contribution models are a total of three CTC models.

The breakdown detector for cluster annotators has been created *by Takamya et al.* [48]**.** Each cluster consists of separate detectors. Results shown in the publications on the proposed CRF baseline system outperformed Japanese DBDC3 tasks.

The empirical assessment of compound indexing for Turkish texts was reported by *Bechikh Ali et al.* [49]. Patterns 9, 10, 11, and 12 are eliminated from studies due to the ambiguity of the pattern. To extract these compound kinds, sophisticated language analysis is required. A considerable improvement has been shown in the Turkish Milliyet IR data set employing compounds for index documents and queries.

Sidorov et al. [23]conducted studies that identify the notion of connecting the information and the context gained by word embedding, which indicates that we may represent phrases. The experiments on STS tasks, i.e., the mean word integration and BoW, are conducted in this work. Experiments provide optimum outcomes for the representation of the sentence. SICK data set is analyzed after operations on STS tasks, which yields excellent performance at a correlation of 1⁄4 0:724. It is excellent since it is hard to quantify similitude with unattended phrase format for semantics on the dataset.

Wang et al. [50] have built many customized service dialogues by using information linked to the history of dialogue and external knowledge production of responses. An application based on external knowledge and dialogue history generating responses provides functional and responsive outcomes compared to the basic technique. However, the lack of background knowledge and timeliness incoherence should be paid greater attention in this suggested system.

The independent language and domain analysis model was introduced by *Tomas Kincl et al.* [51] based on character n-grams. It is tested and exceeds when the combination of domains is taught. The interplay of background noise and speech fluctuations was examined. This research indicates that a Normal Hearing (NH) listener provides more extraordinary performance for stable state noise than a Cochlear Implants (CIs) listener. The performance of both listeners relies on background noise and speaking tempo. However, it is noticed that the speaking rate impacts CI listeners rather than NH listeners considerably. Two algorithms, the SRT and the long-term Mean Root Square, have been employed in this study (RMS).

An English Chinese multi-lingual word representation has been presented by *Yen An-zi et al.* [52]. The author has examined numerous problems relating to word alignment in this

study. It also focuses on multi-lingual word representation. Paper examined several ways of nonalignment and alignment to provide contexts for the Skip-gram model's training. The algorithms employed in this study are cross-lingual semantical word relation, cross-lingual analogy argument, and bi dictionary induction.

The cochlear implant voice perception was proposed by *Fei Chen et al.* [53]. This document outlines the benefits of the bilateral CI hearing. The proposed technique improves the perception of phrases that are segmentally interrupted.

A method to learn distributive vectors of a semantic frame based on a neural network was suggested to *Sangkeun Jung* [54]. The approach suggested used two reconstructive features and an integrated correspondence for a meaningful and valid distributed semantic representation. The System proposed reduces the distance between a semantic frame reader to the semantic vector. Semantic vectors are identical in the vector space.

A half-supervised acoustic model for speech recognition was created by *Emre Yılmaz et al.* [55]**.** This model has been built that gives a voice label. This method also provides speaker labels that lead to the best-automated annotation. This helps to create a more precise and efficient acoustic model. The here constructed acoustic model is multi-lingual.

A framework for four South African language pairings is presented by *Ewald van der Westhuizen et al.* [56]**.** The code-switching approach is used spontaneously and naturally for the development of the ASR system. The swapping of code is growing prevalent because of globalization. Code swapping gives the language model undue reassurance.

The system of voice emotion recognition was evaluated by *Moataz et al.* [57]**.** The last precision was 50% and is expanded to 90%. The algorithms described in this report describe the support vector machines, neural networks, and Gaussian model mixing.

The speech recognition system that transforms Speech into text was created by *Prerana et al.* [45]**.** Two models, depending on the speaker and independent model, are utilized in this work. The computer can convert requests and dictations into text using MFCC and VQ methods. This system is also available. The Mel Frequency Cepstral and Vector Quantization technology are used for the extraction and matching of characteristics.

Simon et al. [46] exhibited different degrees of accommodation that extract the dominance of discussions. This study focuses on the interpretation of language and the impacts of human behavior. For spoken language interpretation, human emotion must be included.

In a robust speech recognition system, *Vincent et al.* [58] evaluated several parameters. This study discusses the environment with three factors, data simulation, and microphone. For automated voice recognition technology, this article employs MVDR and DNN algorithms.

This document presents a voice recognition methodology using the KALMAN FILTER bidirectional non-stationary algorithm for the construction of ASR. A two-way voice-to-text conversion system has been built by Mitsubishi et al. [59]. This is a solid system that is simple and sturdy.

In order to improve noise reduction in the audio signal, *Ismo et al.* [59] utilized improved spectral resolution. "The aim of this study is to eliminate noise in the audio signal Spectral resolution improvement and spectral-domain processing." "Linear (LR) predictions for modeling audio signals, FIR filter, autoregressive (AR) model, Standard Spectral Subtraction (SSS) technique, and the psychoacoustic spectral subtraction technique -used to calculate 10 masking thresholds - are all algorithms in this work". The resolution improvement approach is based on the time domain extrapolation of the signal. With higher spectral resolution, precise noise attenuation with less signal distortion may be accomplished.

The ASR system for the Kannada language was created by *Thimmaraja et al.* [60]. This research uses two primary methods, Voice Activity Detection Spectral Subtraction (SS-VAD) and Zero Crossing Spectrum Speed Mind Square Error (MMSE-SPZC). Developing a spoken query system to access any data implemented in this article is a tough undertaking. Mittal et al. [62] suggested and built the Punjabi language ASR system within diverse acoustic situations. Higher memory space is the main constraint of the context-dependent unbound paradigm. This technology has been built especially for mobile devices.

For conversational communication, *Songfang et al.* [34] presented language models based on Hierarchical Bayesian. In this suggested system, parallel training models enable the

system to handle large-size corpora. Here, the smoothing N-GRAM and Pitman–Y or Processes are two types of language models.

The Mel frequency cepstral coefficient (MFCC) and the low pass Filter Zero Interpolation (Low Pass) is the *Deepali et al.* [61] ASR method for Marathi numerals (LFZI). 1000 words corpus are utilized here and implemented using MATLAB.

For the acquisition of the language characteristics of multimodal fusion, *Darekar and Dhande* [62] have developed an adaptive learning method to the Artificial Network Neural (ANN), which leads to the hybrid PSO-FF method that combines both FF and PSO functionalities. In contrast to the current methods associated with the various performance measures, such as "precision, MCC, sensitivity, FDR, precision, FNR, specificity, FNR, F1 score, and NPV", the performance for the created recognition system is assessed in the Marathi and Standard databases. The proposed methodology was finally demonstrated to be 10.85% higher than current accuracy techniques.

In 2019, the HMM-SPSS was proposed for the Marathi language by *Patil and Lahudkar* [63]**.** Moreover, HMM has developed pathways of speech parameters utilized for synthesis. The collection of 5300 phonetically balanced Marathi phrases was recorded to train context-dependent MMs by seven, nine, and five hidden states. Custom quality standards have demonstrated that HMMs in seven countries can provide high-quality synthesized speech with a minimum of 5 hidden states.

Articulatory Cepstral Coefficients (ACCs) were introduced in 2019 *by Najnin and Banerjee* [64] in the cepstral of the articulatory time-location signal. The equivalency of MFCCs and ACCs and the extreme combined performance and insulation indicated the efficiency of common sound and articulatory signaling approaches. Furthermore, the approach suggested has proved that ACCs gained superior results in identification and phoneme classification on standard data sets over numerous simulations than traditional ones.

The performance of 5 audiovisual fusion model models, such as the fusion model, fusion model, HMM linked, turbo decoders, and the multi-storm HMM, was studied and compared in 2018 by *Abdelaziz* [65]**.** The usage of a typical LVCSR Kaldi DNN recipe

was comprehensively confirmed in three tests: a clean-shape noise test, a composite training, and clean-train-clean test deployment. Moreover, the whole investigation has been done out using the audiovisual corpus NTCD-TIMIT. The investigation allowed a new standard outcome, which contrasts new LVCSR techniques to the AV-ASR, with NTCD-TIMIT, with freely available visual characteristics and 37 loud and clear sound signals.

Tao and Busso [66] proposed a framework in 2018 which uses visual elements to improve performance. A gating layer has been developed using a deep learning method that reduces the effects of loud visual characteristics while keeping just the data necessary. A subset of CRSS-4ENGLISH-14 audiovisual corpora had 61hs of voice for 105 themes collected by various microphones and cameras. Moreover, utilizing the Gaussian Mixture Model or DNN, the frame was compared with the current MMS. Alongside the model, the HMM multi-stream model was compared. The test results have shown that the proposed framework across all configurations is superior to current techniques and shows the strength of the AV-ASR Gating-based Framework.

Semi-supervised self-encoder to enhance recognition of speech emotion was advocated by *Deng et al.* [67]**.** The objective was to get the benefits of unlabeled and labeled information composition. The nearby objective of supervised learning improvised the unsupervised autoencoder. A wide variety of validations were also conducted on four open datasets with different criteria, including the Emotion Challenge database INTERSPEECH 2009. Finally, the assessment findings demonstrated that the strategy indicated achieved typical performance with less data on the challenge and other occupations being labeled while the other ways were excellent.

Shi et al. [68] showed a significant fusion characteristic of speech data in 2019. Initially, the power-law nonlinear function for CFCC acquisition, which assessed the human ear's auditory properties, was used to extract a new function. Speech enhancement technology was created at the front end of feature extraction. In addition, the first-order difference and extracted characteristics were blended to provide the current mixed characteristics. In addition, an energy element called the TEOCC was extracted to produce fusion feature sets

and combined with the a fore mentioned characteristics. The PCA is used to identify and optimize the feature set. The final feature set is employed in single words, language recognition with a short vocabulary, and non-specific people. Finally, a voice recognition model was created using Support Vector Machine to validate a built feature set (SVM) advantage. The validation results have thereby shown the efficiency of the created feature set.

In 2017 Sharma et al. [69] introduced the multi-level Deep Sparse Representation (DSR) to acquire a vocal representation. A thick layer between two sparse layers was employed for a successful application, rather than a number of sparse layers. Furthermore, the recommended DSR system contains significant data in a discrete sparse layer. Therefore, the final feature representation has been achieved after connecting the depictions purchased at the sparse layer. As concatenation results in high dimensional characteristics, PCA was used to reduce the dimension of the obtained feature. Finally, the recommended characteristics proved to be better to standard voice recognition characteristics.

Annu Choudhary et al. [70] suggested to use Hidden Markov Model Toolkit for the identification of the speech for solitary and linked Hindi words (HTK). Hindi words are employed in MFCC-extracted datasets, and the accuracy of the identification system in solitary words is 95% and in related words is 90%. Preeti Saini et al. advocated the use of HTK to be recognized automatically by Hindi[71].

The speech in HMM topology is recognized by isolated words in 10 states, resulting in 96.61%. The automated Bangladesh voice recognition technology was introduced by Md. Akkas Ali. Linear Predictive Coding (LPC) and Gaussian Mixture Models were carried out using feature extraction (GMM). 100 words were captured 1000 times, giving a precision of 84 percent. Malayalam's word identification system was created by Maya Money Kumar et al. [72].

The suggested study was done utilizing HMM on MFCC to extract features on a syllable basis. The Sanskrit voice recognition using HTK was presented by Jitendra Singh Pokhariya and Dr. Sanjay Mathur [73]. Both MFCC and two HMM states have been employed for extraction, producing accuracy between 95.2% and 97.2%. The real-time

speech recognition system for Hindi Words was created by Geeta Nijhawan [74]. In 2014, MFCC extraction utilizing the Linde, Buzo, and Gray (VQLBG) method. To eliminate the quiet, Voice Activity Detector (VAC) was suggested.

Table 2.1: Literature review

Authors	**Techniques**	**Advantages**	**Disadvantages**
Sun, S. et al [36]	Multi-task learning framework	Enhances the acoustic model's performance while reducing the word error rate.	With noisy signals, this won't work.
Maas, A.L et al.[73]	DNN hybrid speech recognition system	Is able to dissect complex recurrent, sequence-discriminative, and HMM-free systems	Due to the lack of a solid foundation, performance has been severely hampered.
Upadhyaya, P et al.[74]	.CD-DNN-HMM(Context-Dependent Deep Neural Network HMMs)	Further optimization of the DNN layer will yield more reliable findings.	Suffers from overfitting problems
Mao, X. and Zhang, Y [75]	Cross-lingual speech recognition system	Enhances the recognition performance even with sparse training data by using this method.	The lack of objective functions and applications on multilingual training have an impact on optimization.

Mitra, V et al.[40]	HCNN (Hybrid convolutional neural network)	For clean, noisy, and channel mismatched situations, it reduces WER (word error rates).	This rule does not apply to languages other than English while doing speech recognition tasks.
Lee, M et al.[41]	DNN-based ensemble structure	Greater efficiency and accuracy in regular reverberant environments.	The acoustic models' design was not taken into account when trying to capture long-term dependencies.
Kamper, H et al.[76]	Segmental Bayesian model	Use frame-level features to make progress.	When working with unsupervised systems, this leads to a high rate of mistakes.
Xue, S et al.[77]	The hybrid NN/HMM speech recognition model gets a new speaker adaption technique.	Reduces the risk of over-fitting by just changing the singular values in the weight matrices.	For bigger DNN models, speaker adaption requires more complexity.

Jacob et al. [78]	Language pre-train bi-directional representation was used	- 80.4% GLUE score was attained, an improvement of 7.6% relative to the previous result. -Squad 1.1 accuracy of 93,2 percent was achieved.	There is no proof that the proposed strategy works for many different tasks. -Not sure if BERT is useful for studying linguistic phenomena.
Matthew et al. [79]	Contextualized word representation was first introduced and then adopted.	As an example, ELMo reduces relative errors by roughly 5-20 percent for NLP tasks such as question answering and textual entailment, as well as semantic role labelling and conference resolution.	-
Jiataogu et al. [80]	Model Agnostic Meta Learning Technique is used	Outperforms when using limited training resource data and multilingual transfer learning.	Not tested for large datasets.
Guillaume Lample et al. [81]	Proposed neural and phrase-based unsupervised model	Comparing the results of unsupervised and semi-supervised models, we got superior results for the low resource language. The combination of PBSMT and	No use of the proposed method for semi-supervised education.

		NMT, according to these findings, yields even greater results.	
Alexis et al. [82]	Introduced probing and downstream tasks	This study examined three different encoders taught in eight different ways to reveal surprising traits of both encoders and training methods. The probing tasks were designed to capture simple linguistic features in phrases.	No multi-tasking performance testing was done for this version because it is only available in the English language.
Jeremy et al. [83]	Proposed Regularization of attention functions by a recurrent neural network using eye-tracking corpora	Significantly improved the detection of abusive language, sentiment analysis and grammatical error detection across three key tasks.	Collecting eye-tracking corpora for Indian or the majority of other languages is a difficult task.
Maria Jeremy et al. [84]	Universal Language Fine Tuning Model (ULMFiT) for All NLP Tasks Proposed	Reduces mistakes by 18-24 percent for six text categorization tasks, outperforming the competition. The model under consideration is free and open source, and it may	There isn't much emphasis on figuring out what knowledge a pretrained language model holds, how it

		be implemented using tools found on the internet.	changes during fine tuning, and what information is needed for various tasks in the proposed work.
Samira et al. [37]	Develop and compare various approaches such as HMM-GMM, DNN, MLLT and CMVN for the Punjabi ASR system.	ASR was implemented in Punjabi and two approaches were tested: HMM-GMM and DNN-GMM. When utilising the DNN-GMM technique, connected and continuous speech improves by 4-5 percent and by 1-3 percent, respectively.	These results were obtained using a limited corpus of Punjabi text.
Virender et al. [54]	A hybrid framework for speech recognition was introduced, based on HMM-k nearest neighbours	96.86 percent accuracy was achieved by combining HMM with KNN.	Need to improve aspects of confused classes' discrimination, and how to reduce the proposed framework's complexity even further without

			impacting accuracy
Jyoti et al. [19]	Utilizing the Kaldi Tool,created an ASR system for Punjabi.	Using n-gram, three models were created: tri1, tri2, and tri3. Models' performance improved from tri1 to tri3, because of enhanced design. The accuracy of MFCC feature extraction is superior to that of PLP feature extraction.	Only tri phone-based voice signals will be supported by the proposed system.

2.3.1 Speech Recognition System Development Journey

During 1920 there was a recognition of language. A toy to identify voice was made of the first machine, i.e. Radio Rex. At the World Exhibition in New York, Bell Labs created a voice synthesizer. However, they eventually dismissed their attempts on the ground that AI was ultimately necessary for success. In the 1950s, researchers researched the basic ideas of phonetic acoustics in order to design systems for ASR. The vowels evaluate the spectrum resonance values of each syllable in most of the systems in 1950. On Bell Labs Davis, Biddulph, and Balashek (1952**)** [85] used the formant frequencies calculated in the vowel regions of each digit to premeditate a single-speaker digit recognition system. 10 syllable recognizers of a single speaker were produced by the RCA Labs, *Olson, and Belarus in 1950* [33]. Forgie and Forgie created a 10-vowel recognition unit at the MIT Lincoln Lab for speaking vowels. Fry and Denes (1959) were attempting, by use of a spectrum analyzer and a pattern match, an acknowledgment recognizer to recognize four vows and nine consonants at the University College in England [9]. In 1960-70, Japanese laboratories entered the area of recognition. Computers are not quick enough and have created H/W as a component of their system for a unique purpose. Nagata reported a Radio Research Laboratory system in Tokyo and as a H/W vowel recognizer. Sakai and Doshita's

work in 1962 at the University of Kyoto, which established a HW recognition system. In 1963 a digital recognition was constructed by Nagata and colleagues at NEC Labs. This led to a lengthy and fruitful program of study. In 1970, isolated word recognition was the focus of study. In large-scale voice recognition, IBM researchers investigated. Researchers launched independent speech recognition studies with speakers at AT&T Bell Labs. Many clustering techniques were utilized in order to identify the number of separate designs necessary for word recognition by the speaker. This study has been modified to make extensive use of approaches for independent motifs of speakers. The Harphy system of Carnegie Mellon University identifies with acceptable precision speech with a vocabulary dimension of 1011 words. It used the finite state network for the first time to decrease computation and efficiently find nearest matching strings. In 1980, the study focused mostly on the recognition of spoken words. In early 1980, Moshey J. Lasry examined letter and digit speech spectrometric and produced a characteristic voice recognition. Technological changes occur in 1980 in particular in HMM speech research from template-based techniques to statistical modelling. The paradigm change was most important when statistical techniques, particularly stochastic HMM.

(*Baker, 1975 and Jelinek, 1976)* were introduced at the start of the 1970s *(Portiz 1988).* This technique still prevails more than 30 years later. Notwithstanding the simplicity of their language models, N-gram has been quite strong. Most practical speech recognition systems nowadays are statistically based, and their findings have been improved in the 1990s. One of the most important technologies produced is the Hidden Markov (HMM) technique in 1980. IBM, the IDA, and the Dragon Systems have comprehended HMM, although in the mid-1980s it was not famous. The other technique, reintroduced in the late 1980s, is neural networks for voice recognition difficulties.

In 1990, the technique was created to recognize patterns. It has historically followed the framework of Bayes but has been modified to reduce the empirical detection error into a problem of optimization. The rationale for this adjustment is that Bayes' theory cannot be used in this context and that the distribution functions for the speech signal cannot be determined appropriately. However, the purpose is not to suit the data but instead to create

recognizers with the least recognition error. Minimum error classification (MCE) and maximum reciprocal information approaches utilized for error minimization are (MMI). These approaches lead to an approach to language recognition based on maximum probability. The proposal is for a weighted HMM method to solve difficulties of robustness and discrimination based on HMM voice recognition. A maximum probability stochastic correspondence strategy was presented in order to reduce the acoustic discord between the provided speech model set and the test utterance. The usage of a neural network as a vector quantizer is the basis of a narrative method for the HMM speech reconnaissance system, an outstanding invention while training a neural network. A number of approaches have been described for predicting a resilient HMM distribution of probability output. The second order of the HMM in comparison with current Viterbi method has been increased by the Viterbi method. DARPA continued throughout the 1990s. The Air Travel Information Service (ATIS) work and the transcription of news broadcast will be the main emphasis (BN). Advances were described in speech acknowledgment and loud speech acknowledgment. Little work has been done in the field of strong voice recognition. A novel strategy for an auditory model was developed for a noisy environment, enabling solid speech detection. Compared to previous models, this technique is computationally efficient. A model-based approach for spectral estimates was created. A Bayesian variable assessment methodology was created in 2000. It is based on the subsequent parameter distribution. Giuseppe Richardi created the technology to tackle ASR adaptive learning problems.

In 2005, modifications were made to the system for performance enhancement for broad vocabulary continuous speech recognition. The five-year Corpus of Japanese Spontaneous (CSJ) project was carried out in Japan. It consists of around 7 million words, which corresponds to 700 hours of speaking. Acoustic modeling, phrase limit detection, pronunciation modeling, acoustic and language model adaption, and automated speech resuming are the approaches employed for this project). Utterance testing, particularly for spontaneous speech, is being researched to further improve speech reconnaissance systems' robustness. They employ multimodal communication when people talk to one another.

When a conversation is held in a loud setting, it enhances the rate of successful transmission of information. The use of visual face information, particularly lip motion, has already been explored in speech recognition and findings reveal that mixing both modes of data deliver greater performance than using audio or visual information, particularly in noisy surroundings.

2.3.2 Review of Speech Recognition System for Indian Languages

Indian researchers have been working on regional languages in India for the past decade. The majority of Indians are not in a position to grasp English, even though ASR-based technologies have evolved that are not helpful to most Indians. In India, there are a total of 23 languages officially spoken languages. As research is underway in the Indian language speech recognition system, it is still in the development stage and is not employed in commercial applications such as home automation, etc. Table 2.2 covers a study on Indian languages using the technique.

Table 2.2: ASR Systems for Indian Languages

Author	**Language**	**Feature Extraction Methodology**	**Model/ Classifier**	**Accuracy**
[60]	Kannada	*SS-VAD and MMSE-SPZC*	GMM	17.01% and 13.18% for noisy and enhanced speech, respectively
[86]	Punjabi	MFCC	CI, CD-Untied CD-Tied, D_DelInterp	CD-Untied gives the highest accuracy with a rate of 81%

[21]	Marathi	MFCC	DWT, LPC& ANN	78%
[61]	Marathi	MFCC &LFZI	SVM	-
[35]	Marathi	MFCC	HMM & GMM	80-90%
[54]	Punjabi	MFCC and GFCC	HMM-GMM & DNN-GMM	4-5% improved performance in DNN-GMM compare to HMM-GMM
[19]	Punjabi	MFCC & PLP	*N*-gram model, MPE	MFCC gives good results compare to PLP
[51]	Tamil	SS–NE, CSD-NE	FCM with EM-GMM	Accuracy improved from 1.2 to 4.4%
[52]	Assamese	MFCC	HMM,VQ and I-vector	90-100%
[23]	Chhattisgarhi	-	HMM, ANN and SVM	99.84 and 94.24% for isolated words recognition systems using ANN and SVM, respectively
[50]	Hindi	MFCC	Genomic HMM, Segmental HMM, Hybrid HMM	-

[49]	Telugu	MFCC-GMM, Prosodic-NNC	GMM	88-77% accuracy
[48]	Hindi	MFCC	VQ-GMM	93% accuracy
[47]	Gujrati	MFCC	HMM	Accuracy 95.9% in lab and 95.1% in noisy environment.
[44]	Kannada	MFCC	MLLT	Maximum, 90.05% accuracy, achieved for Online Mandi Phoneme based Web application
[43]	Urdu	MFCC	Bidirectional LSTM, RNN	WER is 0.68
[42]	Bengali	MFCC	Bidirectional LSTM	98.9% classification accuracy
[87]	Dravidian	MFCC + SDC	ANN	Dravidian language classification accuracy -73.6%,72%, 65.1%, and 68.8% for Kannada, Malayalam Tamil, and Telugu, respectively

2.4 Challenges in Speech Recognition

Although voice recognition development has been underway since 1920, several issues remain unresolved in the speech recognition system. Issues that have an impact on speech accuracy are mentioned below.

a) **Speech type:** Speech style relies on several factors such as voice tone, accents, speaking tempo, voice pitch, and production of phonemes. Voice tone, normal and quiet, maybe yelled. The focus differs from person to location for the same language. The generation of phonemes depends on the language. Speech may be separate words, associated words, continuous or spontaneous words. Tone pitch also sometimes causes voice fluctuations, and ASR systems can hardly understand the words.

b) **Environment:** The surroundings are one of the hardest barriers to the system of speech recognition. Background noise, room acoustics, and channel conditions may represent the environment. These settings add signal noise and noise to the speech.

c) **Speaker features:** Speech variability relies on the speaker's features. The features of the speaker include the age, sex, and articulation variance of the speaker. Articulation variation involves a speaker's mental state, tension, emotions, etc.

d) **Language features:** Developing a separate language ASR system requires distinct methodologies since each language has its own structure. The grammatical suite and phonetic statements in each language are their own.

2.5 Research Gap

- It is found that there is very rare research has been done in Marathi Speech Recognition.
- A minimal dataset is used for developing Speech Recognition System
- The systems developed for Marathi Speech Recognition had not considered various speech characteristics such as different environments, different speech styles, no. of speakers etc.

- The most important thing is no one considered the dialects for developing the System.
- The conventional approach is used for developing Speech Recognition System

2.6 Research Problem

In Maharashtra, though Marathi is the first language and spoken across Maharashtra. It has officially 12 dialects such as Varhadi, Konkani, Puneri, Malvani, etc. The Speech Recognition research work done till date, and the companies which offer Marathi Speech Recognition have not considered these dialects while developing the System. Speech as an interface is a demanding and growing user interface for upcoming products. Unless we consider these dialects to develop the Speech Recognition System, we cannot develop an inclusive Speech interface-based product. Based on the review, it is observed that the research happened till date considered the conventional approach for developing Speech Recognition System. So, we must design and develop a system that considers the various dialects are spoken in the Marathi Language as well as the deep learning approach, which is widely used nowadays.

2.7 Research Objectives

The objectives of the study are,

A. To gather and pre-process Marathi language training data with different speakers, dialects, and environmental conditions for designing a Speech Recognition system.
B. To design an algorithm to extract features for enhancing the performance of a speech recognition system that can deal with the high-density impulse noises in speech signals.
C. To develop an effective training technique using a new hybrid optimization algorithm for the deep learning classifier, which uses classification parameters to improve the performance for classifying data.

D. To model and establish a speech recognition system for the Marathi language that offers precise and efficient extraction of noise and computational efficiencies when dealing with high-density noises.

2.8 Summary

This chapter has presented an analysis of data on the study design and development of a speech recognition system for the Marathi language. Moreover, the chapter provides studies from Approaches for Developing Speech Recognition System, Speech Recognition Systems for Indian Languages, Challenges in Speech Recognition, Research Gap, Research Problem, and Research Objectives. The next chapter of the study will present the Research Methodology of the study.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

This section of chapters consists of the methodologies and tools used to analyze different approaches for developing the Marathi Speech Recognition System. There are various discussions made in this chapter about the considerations of the methodologies and how it works. We have discussed different techniques for Speech feature extraction and pattern recognition in two different subsections.

3.2 Techniques for Feature Extraction

Feature extraction is one of the significant processes in developing a speech recognition system since it is used to extract speech signal characteristics. Four strategies for extraction of features are given and detailed below.

Linear Predictive Coding (LPC): Speech frames are evaluated to generate a voice vector in linear predictive coding. The input speech signal must be pre-emplaced to extract features when voice signals are transformed into spoken frames or blocks, pre-stress functions. Windowing is the next stage after the pre-stresser. Signal disturbance at the end of the frame at the beginning and the end of the frame is minimized in the window. Window frames, which result in LPC analysis, are linked along with the windowing process. LPC coefficients are produced based on the LPC analysis.

Perceptual linear prediction: Hermansky established the linear paradigm of perception. It uses auditory psychophysics to test the speech signal. SR rate changes as the PLP does not receive speech signals from erroneous information. The forecast is linear, although the SR rate is changed. This is preferred in comparison to the LPC since it focuses on human speech impulses.

MFCC: It is the most common technique used for speech signal feature extraction for developing speech recognition system. The Mel Frequency Cepstral Coefficient(MFCC), in contrast to previous extraction techniques or algorithms, MFCC evaluates human voice

signals more precisely[88]. The extraction procedure for the MFCC feature has numerous phases detailed below, and the image figure is shown in Fig. 3.1.

Pre-emphasis — Used as a step in high-frequency energy amplification of the input voice signal. This enables information to be recognized and recognized in these locations in the course of HMM model training.

Windowing — This step trims the input to distinct chunks of time. The result is an N millisecond large window and an M millisecond long offset. A Hamming window is widely used to avoid the influence on the margin of the rectangular window due to sudden fluctuations.

Fourier Discrete Transform - DFT is used to display the signal in the window, which produces a magnitude and the phase of the signal.

Bank of Mel Filter – The resultant DFT spectrum includes data for each frequency; however, human listening is less sensitive for frequencies beyond 1000 Hz. This idea also has a direct impact on ASR systems' performance, so the spectrum is distorted by a logarithmic Mel scale. Equation (1) is used to calculate a Mel frequency. A bank of filters known as triangle filters is designed with filters evenly dispersed under 1000 Hz and logically spaced above 1000 Hz to achieve this effect on the DFT spectrum. Each Mel filter is known as the Mel spectrum for filtering the DFT signal.

$$mel(f) = 1127 \ln(1 + \frac{f}{700}) \quad \ldots(1)$$

Log **–**The Log offers coefficients for the Mel spectrum.

DCT **—**The last step in MFCC is to convert the discreet cosine on the coefficients in the Mel spectrum. DCT is produced using 13th-order mel-cepstral coefficients.

delta MFCC Characteristics — The first and second derivatives of the MFCC coefficients are also generated and utilized to detect speech changes from frame to frame.

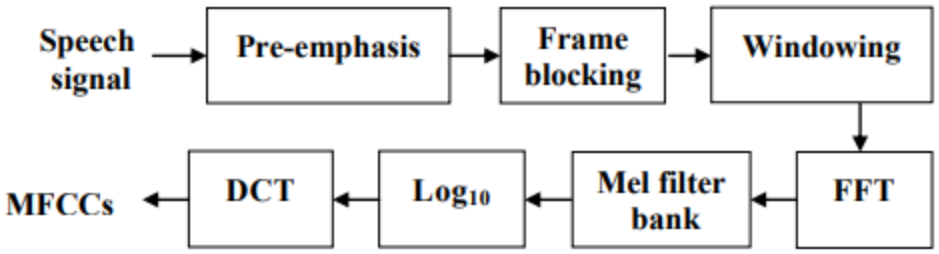


Fig. 3.1: MFCC Feature Extraction

3.3 Machine Learning Techniques

Machine learning is becoming an essential technology to develop various products in firms like Google, Facebook, Amazon, and many more. Computer visualization, voice recognition, advisory system, big data analysis, information retrieval is some of the applications of Machine Learning. Companies that strive to improve themselves apply machine learning approaches to search engines, social networks, movies, music, information, or connect people. Online pages, photos, videos, and sound recordings are classified and marked using machine learning algorithms in web data. Certain items may be found in photos, and a specific music type is detected only when raw data is provided. These approaches allow groups of similar users to be identified, customers' future behavior and items they are interested in based on past customers' data may be recommended.

Machine learning has huge biological and medical applications. Modern biological and medical measuring methods generate an enormous need for novel methods for machine learning. One such methodology is mRNA microarray and sequence method assessment. The measuring data are preprocessed initially, then interest genes are identified, and predictions are created eventually. Additional machines for detecting other splicing, nucleosome locations, gene regulation and so on are employed.

The approaches of machine learning include supervised, unsupervised and semi-supervised. These approaches can be used for solving various problems based on the dataset we have.

3.3.1 Supervised Learning

A supervised machine learning approach is used when we have a sufficient dataset available to train the model. In supervised learning, both the input and output data has previously been submitted to the algorithm. Labeled data is the term used to describe this type of information. Chabot, Robot, and Facial Recognition are some examples of systems that learn through supervised instruction. In this type of data, training data must account for 80 percent of the total data and 20 percent of the test data in order to create a better model for machines. Because the amount of training data grows in large, the amount of corrective systems also grows in size. The three primary forms of supervised systems are once again grouped as follows: regression, classification, and deep learning.

3.3.2 Unsupervised Learning

Unsupervised learning is characterized by the use of unlabeled training data. In its most basic form, this learning strategy is used to discover patterns. It can be used for a variety of tasks, such as discovering hidden patterns and creating appropriate groups. Unsupervised learning can be divided into three types: clustering, dimension reduction, and anomaly detection. Clustering is the most common type of unsupervised learning, followed by dimension reduction and anomaly detection.

Uncontrolled learning is a kind of machine learning that searches for patterns not previously identified in a data set without pre-existing labels and with minimal human monitoring. Uncontrolled training, also known as self-organization, enables the modeling of probability densities by inputs as opposed to supervised learning that typically uses human-labeled data. It is one of the three key machine learning categories, along with supervised and enhanced learning

3.4 Deep Learning

The growth of high-performance computer facilities has made deep learning approaches popular. The ability to handle many features when dealing with unstructured input means that deep learning gains higher power and flexibility. A deep learning algorithm transmits

the data over a number of layers; every layer may gradually extract the characteristics and transfer them to the next layer. Initial layers extract information at a low level, and subsequent layers integrate features to generate a complete representation.

Deep learning refers to the ML families, where the unsupervised functional learning and classification of patterns is carried on via many levels of input processing stages in hierarchical structures. It is situated at the intersections of neural network research, visual modeling, optimization, and signal detection. Today two key reasons are significant; reductions in the cost of computer hardware and the considerably better chip processing capabilities (e.g., GPU units). The efficacy of in-depth learning was established in 2006 in computer vision, phonetic acknowledgment, speech search, the spontaneity of speech recognition, voice and image coding, outer semanticity classification, handwriting recognition, audio processing, and data recovery and robotics. Artificial Intelligence is a broad area where machine learning is a subset of it. Deep learning is a subset of the of machine learning technique. The profound training is a syllable- and photo-understanding method for enhanced machine learning that goes beyond many of its predecessors. It has created enormous successes in extended applications such as speech recognition, picture identification, processing of natural languages, and other industrial goods. Neural networks are used for the creation of machines or intelligent machines.

3.4.1 Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANNs) are motivated to create biologically oriented neural brain networks [16]. However, ANN's are more successful at resolving the difficulties of pattern recognition and matching, grouping and classification, whereas mathematical methods are suitable for linear programming, arithmetical and logic computations.

A relatively rudimentary model of neural connections was used to create the first ANN. The "Mark Perceptron" machine invented by neurobiologist Frank Rosenblatt posits that the artificial connections between neurons may be modified via a controlled process of learning (Rosenblatt, 1957) which decreases them in line with actual and predicted output. A training data set is the expected result. This disparity is spread back throughout the whole

network and enables the weight of the connections to be updated. In other words, the adaptation between the network's current and anticipated responses is the knowledge needed to improve learning performance.

The feed-forward or Layered Perceptron (MLP) models use such technology. The features of these MLPs are three:

- One or more layers of neurons in the hidden layer are not included in the network's input or output layers, which allow the network to learn and solve complicated issues
- The neural activity is differentiated in the nonlinearity and,
- There is a high degree of connectiveness in the network interconnection model

These features and training tackle challenges that are challenging and diversified. Training in a supervised ANN model is sometimes termed an error reverse spread method. The error learning Algorithm is used to create networks based on input-output samples, determines the error signal, the difference between the calculated output and the desired output, and modifies the synaptic weights of the neurons to a greater degree than the resulting error signal and synaptic weight input. Error back learning is based on this concept in two ways:

Future Pass: the network is given with input vectors here. "The input signal propagates forward neurons across the network and appears as the output signal at the output end of the network: y(n) = ć(v(n)) where v(n) is the induced local neuron field defined by v(n) = alternate between w(n)y(n). In the output layer o(n) the result computed is compared to the answer d(n) requested, and the e(n) error is detected for the neuron". The network's synaptic weights stay unchanged throughout this pass.

Backward Pass: The failure signal that was generated on this layer's output neuron is reversed by the network. It estimates the local gradient for the neuron in the individual layers and makes it possible for the synaptic network weights to be changed according to the delta rule as: (n).

This recursive calculation is repeated using the reverse pass for every input pattern, followed by the reverse pass until the network converges. ANN's controlled learning

paradigm is effective and provides answers to many linear and nonlinear issues like classification, plant control, forecasting, forecasting, robotics etc.

Fig. 3.2 depicts how perceptron is given x_1, x_2, ..., x_m inputs (a bias in the inputs is shown in the top left box with the "1" symbol). In addition to the weight gains, the net input function is calculated. In return, the activation function (here is the Unit Step function) is provided, which provides a binary –1 or +1 output that corresponds to the sample's anticipated class label (in this Binary Classification example). During the learning phase, the output itself is utilized to determine the prediction error. An upgrade of the weights backs this to reduce the malfunction between the current and required output.

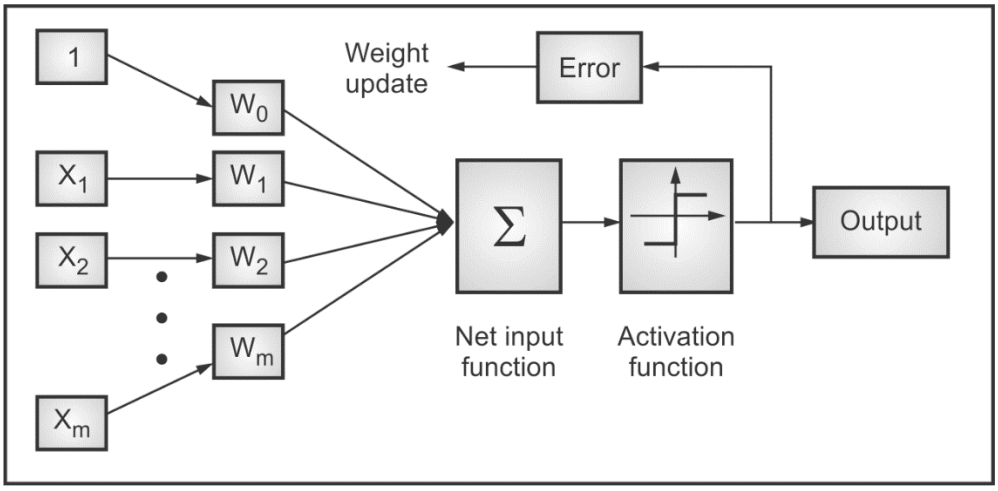


Fig. 3.2: The Perceptron workflow. After Raschka and Mirjalili, 2017 (modified).

About the same time in 1958, the notion of explaining the activity of neurons with respect to digital units was first up by John von Neuman: each neuron might be 'On' or 'Off.' It delivers impulses to neurons that form a nerve aggregate when activated. According to Neuman's method, the activity of neurons would be represented entirely by binary behavior: an ensemble of binary functions that run in parallel may codify the activity of a whole neuronal population.

Bernard Staying connected, and Marcian Edwin Hoff at Stanford University made a stride advance in 1960, as we shall see. They designed two neural network systems: ADALINE

and Jurisdiction (Multiple Adaptive Linear Neuron). The first practical achievements in digital logic and noise deletion for telephone calls were obtained utilizing ANN ideas.

Several achievements and disappointments defined the history up until 1985 of the ANN. A large portion of the science community has been critical of ANNs (Fodor and Pylyshyn, 1988). Cognition in Geosciences summarized that critical view (2013, paragraph 1.2.6: Critiques to Connectionism). It begins with "... two fundamentally divergent intelligence views articulated by the so-called computerists" and "connectionists." The key distinction between classical computer systems and neural is that the first system type is inherently sequential, whereas an artificial neural network functions with a parallel architecture. This distinction profoundly influences the understanding and is the foundation for the other two approaches to studying the human mind. The intelligence itself remains in the weight of the links between the network components in connectionist systems. This implies that the machine can learn the strength of the links by iteratively updating them. The computerists, on the other hand, argue that the human mind operates by applying continuous calculations. One of the most severe opponents of connectionism was Jerry Fodor. The cornerstone of the whole approach to artificial intelligence and not simply the neural networks are criticized by his 1975 book The Language of Purpose. According to Fodor, this modeling and simulation-based method is improper, just as the effort to duplicate the physical world was unfit to study. The tool for building a machine like the actual world is not classical physics, for instance. Although criticism from Fodor and other computer scientists is helpful to underline the limits of connectionism and neural networks, it is inaccurate and improper to see that the human mind functions as a serial computing system." (Christian Friends, 2013).

Marvin Minsky, founding member of the MIT AI Lab, and Seymour Papert, then lab director, were among the harshest criticisms of ANNs. In their work, the authors emphasized the limitations of a Perceptron, such as its incapacity to learn the basic Boolean XOR function, as it cannot be linearly separated. They have rigorously analyzed Perceptron’s' limits and published its conclusions in an important book (Minsky & Papert,

1969). A lengthy period of disenchantment with ANNs is usually thought to generate and encourage this publication, which freezes the financing and publications.

3.4.2 Deep Belief Networks (DBN)

Professor Hinton proposes Deep Belief Networks in order to overcome the limitation of previous neural networks[89]. Two kinds of neural networks – Belief Networks and Restricted Boltzmann Machines (RBNs)- comprise Deep Belief Networks (DBNs). DBN is an unsupervised learning technique in contrast with perceptron and background neural networks.

a) Belief Network

A belief network consists of layers of stochastically weighted binary units. The network is also an acyclic chart which enables us to see in the leaf nodes what kind of data the believing network believes in. A belief network aims to determine the state of the stochastic binaries that have not been seen and to modify the weights between them such that the network can create data similar to that seen. The binary units stochastic in creed networks have a 0 or a 1 status, and a weighted input from other units determines the likelihood of 1. The equation of probability for these units is shown in the Fig. 3.3.

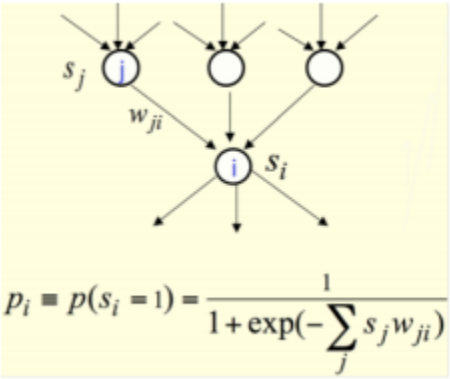


Fig. 3.3: The structure of the belief network and its probability equation

The challenge in learning weights in faith networks is to get a post-distribution that discloses the problem. When two independent hidden units 'explain away,' they might become dependent if they impact both. The occurrence of any of the hidden units may

explain the occurrence of the unit-linked from both hidden units. It is also termed conditional dependency. Moreover, if belief networks contain inter neural networks, the subsequent distribution relies on the previous and the probability of concealed higher levels, and there are different ways in which these layers may be configured. Hinton thus offers a concept to learn from one layer at a time and limit the interconnection of stochastic binary units to make learning fast and straightforward.

b) Restricted Boltzmann Machine

Boltzmann Machine is a recurrent stochastic neural network with random binary units and undirected edges. Unfortunately, Boltzmann's learning is unworkable and has a problem with scalability. This has resulted in the introduction of the Restricted Boltzmann Machine (RBM), which contains one layer of hidden devices and limits connections between hidden devices. This makes learning algorithms more efficient. In Fig. 3.4, the structure of RBM is shown:

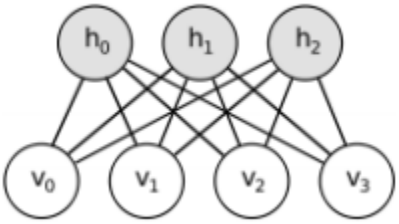


Fig. 3.4: The structure of RBM

Given these setups, the energy function defines probability distributions across hidden and/or visible units:

$$P(v,h) = \frac{1}{Z}\exp\big(-E(v,h)\big) \qquad \textbf{…(2)}$$

Where Z:

$$Z = \sum_{v,h} \exp\left(-E(v,h)\right) \qquad \textbf{…(3)}$$

Then the highest probability learning algorithm can form the network simply by simultaneously updating all the hidden units and all the visible units:

$$\frac{\partial \log P(v)}{\partial w_{ij}} = \langle v_i h_j \rangle_0 - \langle v_i h_j \rangle_\infty \quad \text{…(4)}$$

The principle of consolidating the RBM learning consists of updating, beginning with visible units, visible units from the hidden units, and lastly updating the hidden units, all hidden units in simultaneously. The rule of study is:

$$\Delta w_{ij} = \langle v_i h_j \rangle_0 - \langle v_i h_j \rangle_1 \quad \text{…(5)}$$

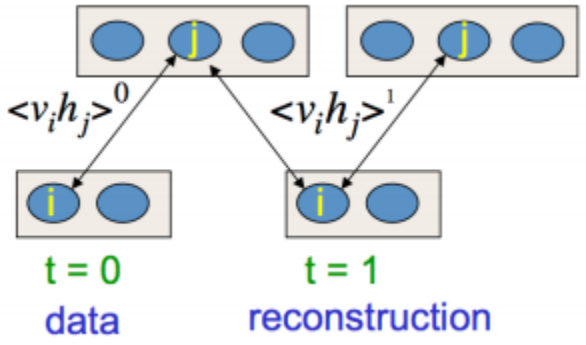


Fig. 3.5: Contrastive Divergence algorithm for RBM

c) Deep Belief Network

Restricted Boltzmann's Machine (RBM) is each layer and is layered to build DBN. Initially, DBN is trained by means of a CD algorithm to learn a layer of characteristics from visible devices.

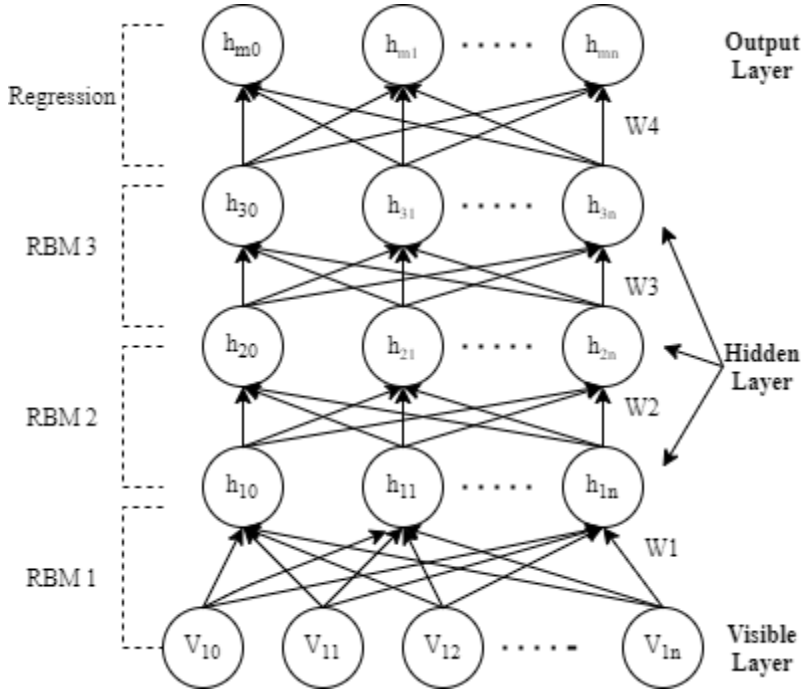


Fig. 3.6: DBN Architecture

In the following stage, the activations of previously taught features are treated as visible units, and features of a second, hidden layer are learned. Finally, when the last hidden layer is learned, the full DBN is trained.

The RBM training using CD algorithms seek ideal local conditions for each layer, and the following stacked RBM layer takes the best learned values and seeks the best location again. In the end, it is probable that each layer will be trained to acquire the best global result at the completion of this approach.

3.4.3 Deep Neural Networks (DNNs)

The basis for many contemporary Artificial Intelligence (AI) applications is the current Deep Neural Networks (DNNs). The number of applications using DNNs has expanded since the initial application of DNNs to voice recognition and image recognition[90]. These DNNs are used in a variety of applications, from cancer detections to sophisticated games. DNNs are currently capable of exceeding human precision in several of these fields. The improved performance for DNN is due to their capacity after utilizing a considerable

amount of statistical learning to extract high-level properties from raw sensory input to generate an effective result.

3.4.4 Recurrent Neural Network (RNN)

A Recurrent Neural Network (RNN) is essentially characterized by having at least one feedback link so that activations may run in one loop round the network[3]. This makes temporary processing possible and allows networks to learn sequences, such as sequence validation or temporal association. The current architecture of the neural network may be multiform in form. The conventional Multi-Layer Perceptron (MLP) with extra loops is one frequent variant. This may take use of the strong non-linear mapping functions of the MLP and have a certain memory. Others have more uniform architectures and might have stochastic activation functions, with every neuron connected to all the others.

RNN is the neural network type where outputs of the previous stage are provided as input into the present phase. A current network is a form of neural network. In conventional neural networks, all inputs and outputs are independent, but in situations such as when a phrase has to be anticipated, previous words are required, and preceding words must be recalled. This led to the development of RNN that solved this issue with the help of a hidden layer. The hidden condition that remembers specific sequence data is the main and most important feature of RNN.

The RNN has a "memory" that recalls all computed information. Each input employs the same settings since all inputs or hidden layers do the same work for the output. In contrast to other neural networks, this minimizes the complexity of parameters.

To correctly grasp RNNs, we need a solid understanding of "regular" neural and sequential data feedback networks. Sequence data is essentially organized data that follow one another in connection. E.g., financial or DNA sequencing data. Perhaps the most frequent sequential form is time-series data, which is a succession of time-listed data points.

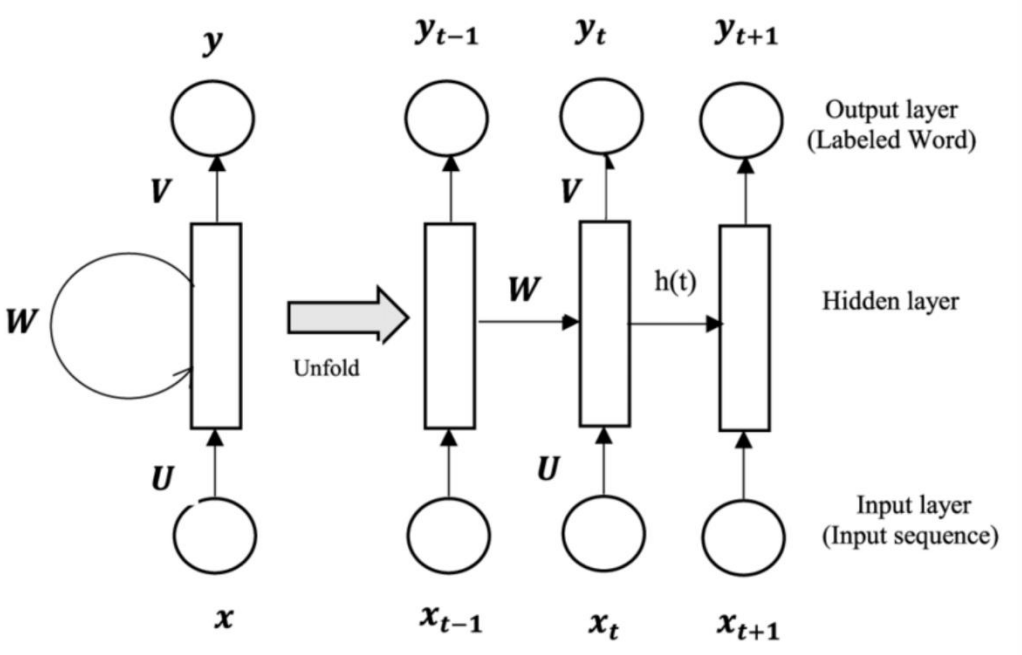


Fig. 3.7: Recurrent Neural Network (RNN)

RNN's and neural feed networks derive names from the way information is sent.

The information in a feed-forward network goes from the layer of input, via the hidden levels, to the output layer in only one way. The information passes directly across the network and does not affect a node twice.

Neural feed networks have no recall of the information they receive and cannot predict the next. Since a feed-in network takes the current input into account, it has no sense of a timely order. It can simply recall nothing, save schooling, about what occurred in the past.

When taking a decision, it takes account of the current input and the inputs it has learned. The following two pictures highlight the difference in the flow of information between an RNN and a neural feed system.

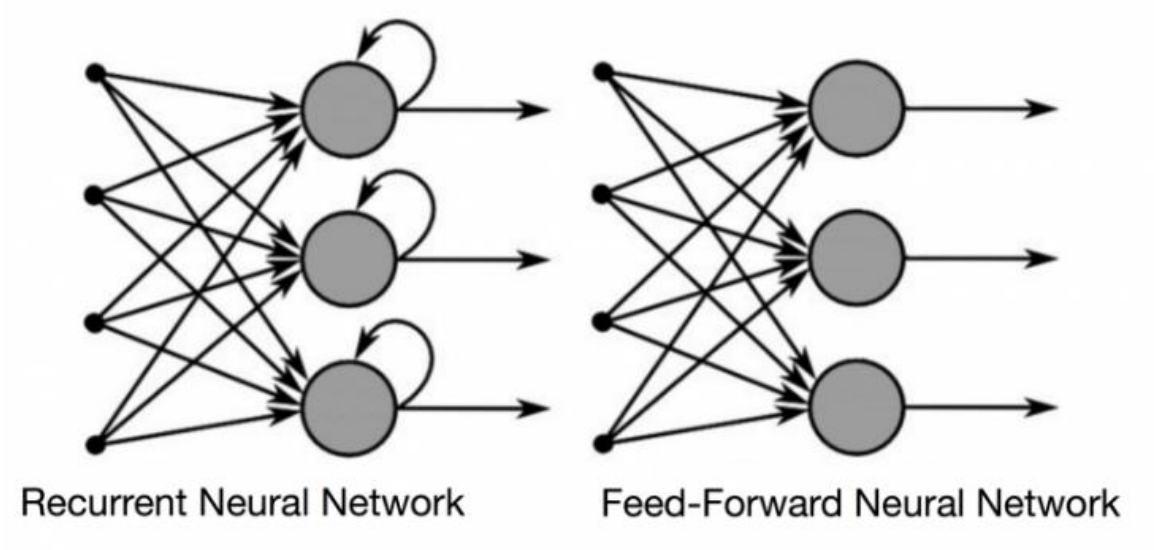


Fig. 3.8: RNN and feed-forward network

An ordinary RNN has a short memory. They do have a long-term memory in tandem with an LSTM (more on that later). The notion of a recurrent neural network memory may also be shown nicely to explain with one example. Let's imagine that you have a regular neural network feed-forward and give it the word "neuron" as the input and analyze the character of the word. When the letter "r" is reached, it has already forgotten "n," "e," and "u," which make it almost difficult to forecast which character will follow for this sort of neural network.

However, because of its internal memory, a recurring neural network may recall these characters. It generates output, replicates it, and retrieves it into the network. In short: recurring neural networks bring to the present an immediate history. Therefore, the now and the recent past have two inputs to an RNN. The data sequence provides meaningful data about what is next; therefore, an RNN can accomplish things other algorithms don't do.

As with all other deep-learning algorithms, a feed-forward neural network gives a weight matrix to its inputs and generates output. Note, for the present as well as for the prior entrance, RNNs apply weights. In addition, a recurring neural network also changes weights both through gradient descent and time spread (BPTT).

Note that whereas neural network feeding charts one entry to one output, RNNs may map one to many, many to many, and many to one output (classifying a voice)**.**

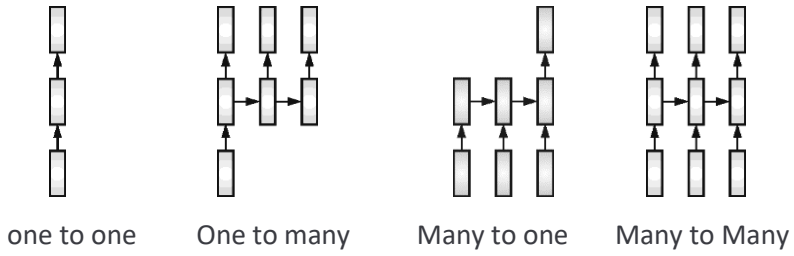


Fig. 3.9: RNN Models

Backpropagation

You will first have to comprehend the notions of forward and back propagation in order to comprehend the idea of backpropagation across time. We could pass a whole essay on these topics therefore I'll try to provide a definition as basic as feasible.

You essentially spread your model's output in neural networks to see whether it's true or wrong to obtain the error. Backpropagation is nothing more than traveling backwards through the neural network to determine the partial derivatives of the weights mistake, allowing you to remove this value from weights.

These derivatives will subsequently be employed using gradient descent, a process that may reduce a given function repeatedly. After then, the weights are adjusted up or down to which the inaccuracy diminishes. This is how a neural network learns throughout the training. We essentially attempt to adjust the weight of your model while exercising with backpropagation.

The Fig. 3.10 shows the notion of future propagation and propagation in a neural feed network:

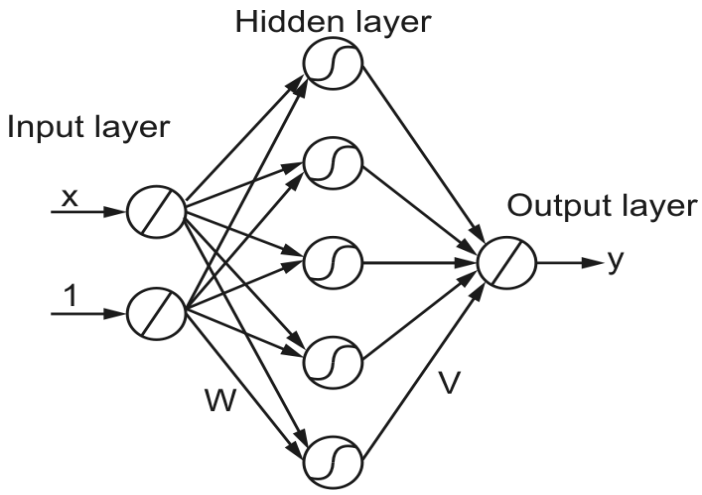


Fig. 3.10: Feed-forward network

Backpropagation through time (BPTT) is a fantastic buzz phrase for unrolled RNN reverse propagation. Unrolling is a conceptual visualization tool that enables you to comprehend what's going on within the network.

The following graphic shows an unrolling RNN. Left, following the same indication, the RNN is unrolled. Note that there is no cycle after the same sign since the various time steps are shown and information is transmitted from one step to another. The Fig. 3.11 also explains why an RNN may be seen as a neural network sequence.

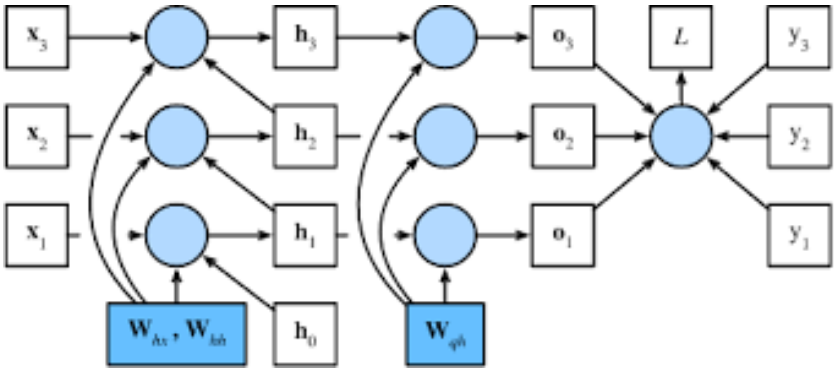


Fig. 3.11: Back-propagation in RNN

If you perform BPTT, you will need to conceptualize the unrolling since a time-step mistake relies on the preceding step.

In BPTT, the fault is propagated back to the initial step, and all times are unrolling. The result is that the inaccuracy may be calculated at any moment, allowing weights to be updated. Note that if you have a large number of steps, BPTT may be very costly.

Challenges in RNN

There are two key challenges that RNN has had to face, but you first need to know what the gradient is to comprehend it.

In terms of its inputs, a gradient is partially derivative. If you don't know what it implies, simply think about this: if you modify the entries a little, a gradient estimate just how much output of a function will vary.

You may also see the slope of a variable as a gradient. The greater the pitch, the steeper the pitch, and the more quickly the pattern can learn. However, the model quits learning when the pitch is zero. In respect of the error change, a gradient quantifies merely the change in all weights.

a) Explosive gradients

Explosive gradients occur when weights are given foolish significance by the algorithm without much rationale. Thankfully, truncating or squeezing down the gradients may readily address this issue.

b) Gradient Descent

Declining gradients occur when the gradient values are too tiny, and the model stops learning or takes too much time. This was a big concern in the 1990s and much more difficult to fix than the explosive gradients. Luckily, Sepp Hochreiter and Juergen Schmidhuber solved it with the LSTM idea.

3.4.5 Long Short Term Memory (LSTM)

The LSTM Networks are one of the most advanced designs for profound learning, including handwriting, voice recognition, and series forecasting (Hochreiter and Schmidhuber, 1997; Graves et al., 2009, 2013)[91][92]. It is surprising because we know that no previous attempt was made utilizing LSTM networks to evaluate performance in a big, liquid, and survivor-free stock universe by forecasting large-scale financial market

jobs. Selected applications focused on S&P 500 predicted volatility, selected foreign exchange rates (Giles et al., 2001), and news inclusion for selected companies as in Xiong et al (2015)[93].

Long, short-term memory networks extend the memory for deep neural networks. Therefore, it is advisable to take advantage of important events which entail major delays. LSTM units should be used as building units for the RNN layers, often called the LSTM network. LSTMs allow RNNs to retain inputs for a long time. The reason is that LSTMs have memory information, much like a computer's memory. The LSTM is capable of reading, writing, and deleting memories.

This memory may be seen as a gated cell with a gated meaning, and the cell determines that information is to be stored or deleted (i.e., if the doors are opened). The weights that are also learned by the algorithm give significance. This implies that over time it learns what information is and what is not essential.

There are three gateways to an LSTM: enter, forget, and exit. These gates regulate whether fresh input may be allowed (input gate) to erase the information because it does not matter (lose gate) (output gate). An example of an RNN with its three gates is shown in Fig. 3.12.

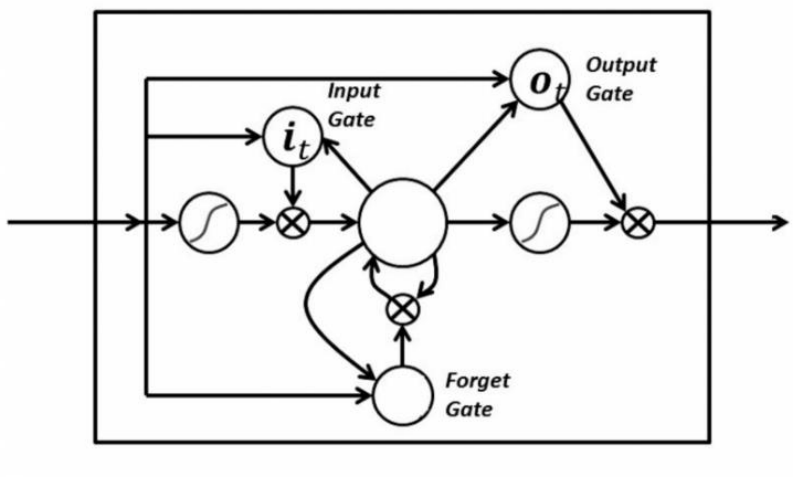


Fig. 3.12: Long Short Term Memory Network

The LSTM gates are analogous to sigmoids, which means they vary from 0 to 1. It is analog to them that allows them to reproduce. LSTM resolves the problem with the extinction

gradients since it maintains the gradients steep enough to maintain a reasonably short training and great accuracy.

3.4.6 Convolutional Neural Network (CNN)

From image processing to speech recognition, the Convolutional Neural Network has achieved groundbreaking breakthroughs in the last ten years in many disciplines of pattern identification. The most advantageous element of CNNs is to reduce the number of ANN parameters. This success has led academics and developers alike to approach bigger models to handle complicated problems that standard ANNs could not do; The primary assumption of issues handled by CNN should not contain spatially dependent properties. For example, we do not have to be careful where the faces are situated in the photographs using a Face Detection program. The key thing is to discover them irrespective of their location in the pictures. A second crucial component of CNN, as the entry propagates to further layers, is to get abstracted characteristics.

A Convolutional Neural Network (CNN) is a deep learning method that can capture input images, attribute significance to distinct elements and objects of the picture (learnable weights and biases), and discriminate between them. In contrast to other classification techniques, preprocessing is substantially lower in a ConvNet. Whilst handmade filters are produced in rudimentary approaches, ConvNets can learn these filters/characteristics with adequate training.

Convolutional neural networks consist of many artificial neuron layers. Artificial neurons are mathematical functions which compute the weighted total of various inputs and provide an activation value.

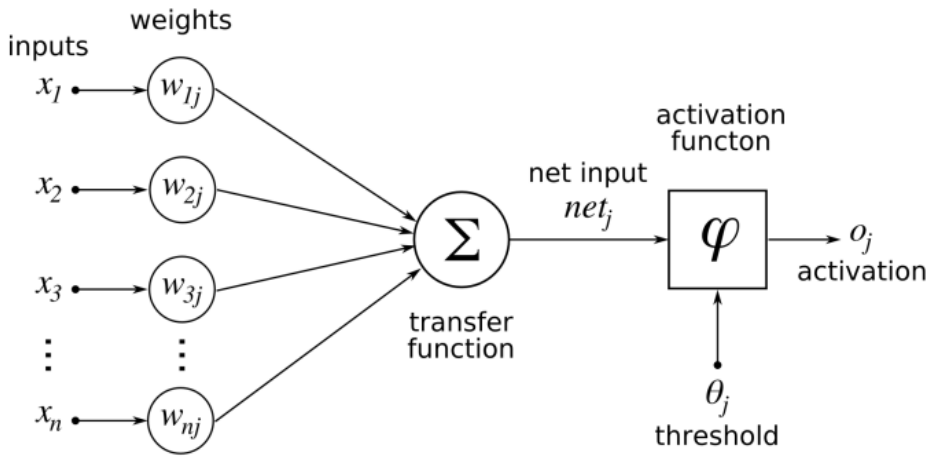


Fig. 3.13: Convolution Neural Network

A weight defines the behavior of each neuron. The artificial neurons of a CNN choose distinct visual characteristics when fed with the pixel values.

Each layer provides many maps for activation when you enter a picture to a ConvNet. Activation maps emphasize the corresponding visual characteristics. Each neuron accepts a patch of pixels as its input, multiplies the weights of their color values, summarizes them, and performs the activation function.

The first layer (or bottom) of the CNN generally identifies fundamental functions such as the horizontal, vertical, and diagonal borders. The first layer output is given as the following layer entry, which extracts more complicated characteristics, such as curves and edge combinations. As you enter the neural network further, layers begin to recognize greater levels of characteristics, for example, objects, faces, and more.

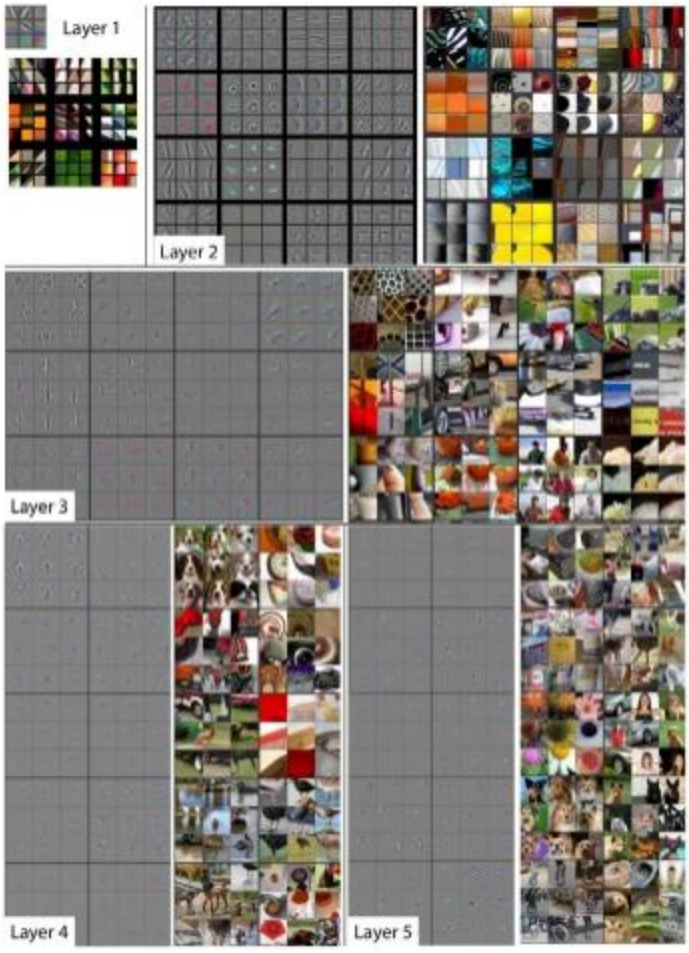


Fig. 3.14: Each layer of the neural network will extract specific features from the input image. (source: http://www.deeplearningbook.org)

"Convolution" is the functionality of the multiplication and summation of pixel data (hence the name convolutional neural network). A CNN usually includes many convolution layers, although it does also have additional components. The last layer of a CNN is a classification layer, which receives the previous convolution layer's output (remember, the higher convolution layers detect complex objects).

Classifications are based on the activating map of the final convolution layer and show how likely the image is to be "class" with a number of confidence rating (such as between 0 and 1). If you have ConvNet, for example, which identifies cats, dogs and horses, the output of the end layer may contain one of the animals.

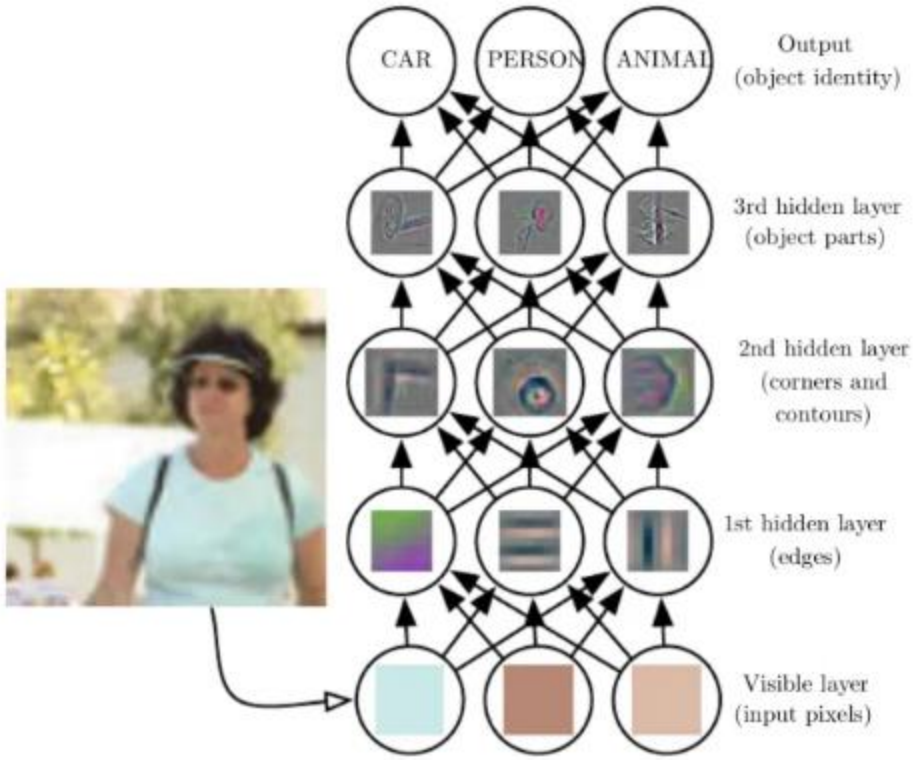


Fig. 3.15: The top layer of the CNN selects the picture class based on characteristics collected by convolutionary layers (source: http://www.deeplearningbook.org)

3.5 Speech Corpus

We obtained permission from our University to obtain this corpus of speech from Indian Language Proliferation and Development Centre, Government of India. According to our request, the Government of India contributed this corpus of speech for the purpose of the research work. There are 44521 wave sample files in the speaker corpus. This corpus of speeches consists of 1000 speakers from the various Maharashtra districts. The frequency of the speech file is 16 kHz. The speakers are captured in various environments. This corpus also contains the utterances of each file which helps us in modeling the recognition system.

3.6 Summary

This chapter fully explains the research methodologies utilized for the study. The study has shown the tools used to beautiful in the study, and the best approach picked for the current investigation. In addition, the researcher provided thorough advice on selecting suitable study instruments. The speech corpus is collected from the Indian Language Proliferation and Development Technology Center, Government of India.

CHAPTER 4: DEEP BELIEF NETWORK USING DUO FEATURES WITH HYBRID-META-HEURISTIC APPROACH FOR DEVELOPING MARATHI SPEECH RECOGNITION SYSTEM

4.1 Introduction

Many scholars, including Marathi, Punjabi, Kannada, and others, have recently explored the recognition method. Many studies have shown that noise is a serious concern when designing a speech/voice recognition system. People who work in extremely demanding auditory environments are proficient at language recognition. Current methods for developing speech recognition systems act in a very circumstantial manner; however, they are not trustworthy or competent to recognize human speech in environments where voice signals are noisy. Bi-modal (audiovisual) is a current or non-conventional technique for developing and testing an ASR system in a noisy environment[66]. Lip readings improve the device's speech recognition capabilities by rectifying distorted voice signals. The lip-reading approach has been used in several studies to increase the performance of SRS systems in video speaker recordings and noisy environments.

The Marathi Language is the official language of the Indian state of Maharashtra, and it is written in Devnagari using the Bramhi script. Over 90 million people globally speak Marathi[17]. However, due to the differences in each language, there is an excellent potential for developing speech recognition systems for all Indian languages. The intensity required to construct ASR systems in Indian languages such that industry or individuals can use them in real-time systems or communication is not critical. This effort has so been made to focus on the Marathi language. This study aims to apply and analyse various classification algorithms for the development of a Marathi speech recognition system. Because of the various dialects in the Marathi language, this work may be critical. It is said that after 12 kilometers, the Marathi dialect changes. In general, spoken signal processing is a difficult task due to numerous hurdles, yet outstanding research can resolve these

concerns. or any other language spoken in India ASR is in high demand in India due to the country's digitalization. Various supervised and unsupervised learning algorithms for various applications in machine learning have been developed throughout the last decade. The majority of these solutions use unlabeled models, such as deep Boltzmann machines with unlabeled data. Autoencoders and other similar devices

The Speech Recognition System is a natural language processing application. Natural language processing is used by the Speech Recognition System. Due to the rapid growth of technology and its application, it is becoming more popular and accepted to its full extent. You can construct Speech Recognition Systems with different approaches, such as the traditional HMM-GMM technique, Deep Learning, or a hybrid strategy.. Depending on the situation, each technique has benefits and drawbacks to consider. The majority of the world's languages benefit from having a big speech corpus available for use in developing a speech recognition system. On the other hand, when developing a speech recognition system, languages that only have a small geographic presence or are spoken by a small percentage of the global population are given less consideration. It's estimated that there are 42 distinct Marathi dialects spoken throughout Maharashtra and the rest of India. There hasn't been much work done to date on developing a Marathi voice recognition system.

In this work, various pattern recognition techniques are compared to find the best suitable for Marathi language SRS development. Firstly, the preprocessing has been done for the Marathi speech corpus, and after that, the feature extraction is implemented using the MFCC algorithm. The process applied here is shown in Fig. 4.1

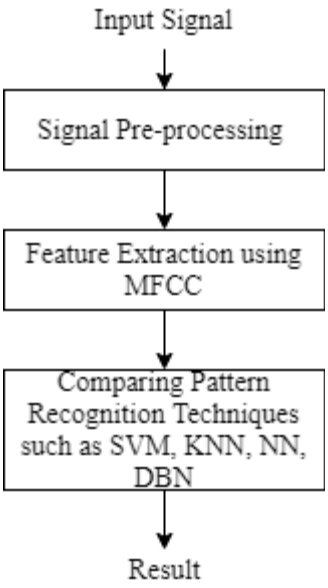


Fig. 4.1:Pattern Recognition Methodology Comparison

After feature extraction, different pattern recognition approaches are executed to find the best suitable one. After performing this experiment, it is observed that DBN is performing well compared to the conventional machine learning methodologies, and results of the same are discussed in the result section of this chapter.

A deep belief network is a two-layer stack of Restricted Boltzmann Machines (RBMs)[94]. There are two types of RBM layers: the visible and the hidden. When compared to deep neural networks, the RBM architecture learns all of the input at once rather than layer by layer. RBMs are undirected energy-based models having two layers of visible and hidden units, with just connections between the layers. An RBM is one of these models. A hidden layer bias for forward passes exists in RBM, as do visible layer biases for backward passes. RBM has two biases. A contrastive divergence strategy is used to train each RBM module one at a time, unsupervised. For example, regression, feature learning, and dimensionality reduction are all possible uses for RBM.

The number of hidden units, layers, and iterations all affect DBN speed. DBN's performance degrades as the number of concealed units increases. As the number of hidden layers increases, DBN's performance degrades. Increase the number of RBM training

iterations to boost DBN's performance. Formula (1) represents the entire RBM energy with the concealed layer (v)

$$\Delta BZE(c_a) = \sum_{a}^{b} ns_a B_{a,b} + \beta_a \qquad (1)$$

The neural network consists of three layers of neurons with a hidden layer and an output layer[40]. A neural network is a feedback network and can only be one hidden level or numerous hidden layers. It has been trained and reproduced. It is being monitored. If the neural network has several hidden layers, it becomes problematic since it must be taught at the same time. There are numerous layers in the neural network that are always hidden. First, controlled learning is employed for training and input reproduction. It is then retrained with a supervised technique of fine-tuning and grading. The DBN hidden layers, i.e., RBM layers, were trained on each layer. It takes longer, but this gives better precision and reduces WER compared to other methodologies.

4.1.1 Related Work

In 2018, Darekar & Dhande [62] developed an adaptive methodology for learning the linguistic characteristics of multimodal fusion by the Artificial Neural Network (ANN) which led to a hybrid PSO-FF algorithm that combines the Particle Swarm Optimization (PSO) and the Firefly (FF) functions of the network. In contrast to existing methods, performances were assessed using several metrics, including “accuracy, mcc, sensitivity, FDR, precision, FNR, special characteristics FNR, F1 Score, and NPV,” in combination with Marathi and standard data points. Finally, 10.85% more precisely than the existing methods have been shown.

Patil and Lahudkar [63] proposed the HMM-SPSS for the Marathi language in 2019. Furthermore, the HMM was used to generate trajectories of speech characteristics for synthesis. A set of 5300 phonetically balanced Marathi phrases was recorded in order to train the context-dependent HMM in seven, nine, or five cached states. Individualized quality metrics revealed that HMMs with 7 States could provide an excellent quality synthesised speech compared to five hidden states with the shortest time.

Najnin and Banerjee[64] implemented Articulator Campestral Coefficients (ACCs), which were the cepstral coefficients of articulate time-location signals, in 2019. Furthermore, the proposed methodology demonstrated that ACCs outperformed existing methods in recognizing and categorizing phonemes on standard datasets over many simulations. The MFCC and ACC equivalence, as well as the remarkable combined performance and isolation, demonstrated the effectiveness of the common methods for sound and articulatory signals.

The audio-visual fusion models in 2018, including the Decision fusion model, feather model, HMM coupling, turbo decode, and multi-stream HMM, were studied and compared by Abdelaziz [65]. The typical DNN-based LVCSR Kaldi recipe was verified in three testing methods comprising a clean-train-noisy trial, combined training, and clean-train-clean-test deployment. In addition, the NTCD-TIMIT audiovisual corpus has been thoroughly analyzed. For this investigation, the visual characteristics of NTCD-TIMIT and 37 loud, clear, and publicly accessible sound samples were utilized to evaluate new LVCSR AV-ASR methods with a conventional result.

In 2018, Tao and Busso suggested a performance improvement approach utilizing visual characteristics. Profound learning provided via a grading layer reduces the effect of noisy visual characteristics and only maintains the important data[66]. The framework was constructed using an audio-visual subset of CRSS-4ENGLISH-14, which consisted of 61 hours of 105 speakers and therefore was collected from multiple microphones and cameras. In addition, a comparison was drawn between the framework and conventional HMMs utilizing the Gaussian Mixture and DNN monitoring system. The model was also compared with the HMM multi-stream model. The test results showed that the suggested framework exceeded existing methods across all configurations and showed the power of the AV-ASR Gating Framework.

Deng et al. (2018) recommended semi-monitored self-encoders to improve recognition of linguistic emotion[67]. The goal was to reap the benefits of unlabeled and labeled information composition. The goal of supervised learning was to improve an unattended auto-encoder. Furthermore, we performed a wide range of validations on four publicly

available datasets and the INTERSPEECH 2009 Emotion Challenge database. Finally, the assessment findings reveal that the suggested methodology achieves normal performance with less labeled data on the challenge and other tasks and outperforms the other current methods.

Shi et al. [68] published a robust fusion feature that categorized speech data as its whole in 2019. The power-law non-linear feature was extracted first in order to obtain CFCC, which evaluated the auditory qualities of the human ear. Speech enhancement technology was developed at the beginning of the function extraction process. Furthermore, the first-order distinction and extracted feature were blended to create the most recent mixed functions. In addition, a fusion feature set was retrieved and blended with the previously indicated features. The collection of features was discovered and optimised using principal component analysis (PCA). The final set of attributes was used in single words, small-speak languages, and non-specific people. Finally, a voice recognition model based on Support Vector Machine was developed to validate the benefits of the established feature set (SVM). As a result of the validation findings, the feature set developed is effective.

Sharma et al.[69] provided the multi-layered Deep Sparse Representation (DSR) in 2017, allowing the acquisition of a feature for speech recognition. Instead of a sequence of sparse layers, a thick layer between two sparse layers is employed to encourage effective utilization. Furthermore, the recommended DSR system has a notably sparse layer containing valuable data. As a result, a final feature presentation was achieved after connecting the representations obtained in a sparse layer. Because concatenation produces high-dimensional characteristics, PCA was employed to reduce the acquired characteristic dimension. Finally, it has been demonstrated that the recommended attributes outperform standard speech recognition features.

“Today's world is data-driven, and we must focus on approaches to increase human living standards. Natural language processing and comprehension are critical in converting technology on a different level. If we want to improve their lives through NLP, we need to work on regional languages such as Marathi, Hindi, Punjabi, and so on.”

Supriya et al.[95] developed HMM and DNN-based Marathi Speech Recognition System. 1500 speakers are represented in the dataset used in this implementation, totaling 15000 voice recordings. The Kaldi ASR toolbox was used to implement the suggested system, and a Word Error Rate of 24% was achieved. 340 sentences of text were used while performing. Five models were compared for the comparison analysis, with SGMM outperforming the others (Mono, Tri1, Tri2, Tri3, and SGMM). Kishori et al. [96] proposed a Marathi ASR system based on neural networks for single words. There are three utterances of each word in the voice corpus used for implementation; the DWT technique is used in this paper for feature extraction. Pattern identification was aided by an Artificial Neural Network (ANN), which was 60 percent accurate in its results. S. Lokesh et al. [97] created a Tamil Speech Recognition System using a Bi-directional Recurrent Neural Network. Pre-processing voice data using the SGF algorithm. For feature extraction, the MAR and PLP algorithms were used, and the classifier was a BRNN. BRNN-SOM, RNN, and DNN-HMM classification algorithms were all implemented and studied to see how well they worked. The proposed technique outperformed existing classifiers in terms of accuracy, coming in at 93.6%.

Sangramsing used the HTK toolkit to build Marathi Speech Recognition. In this case, MFCC is employed for feature extraction, while HMM is used for classification[98]. Thus, the word error rate is 5.37%, while the accuracy rate is 94.6% Because of this, only 30 words from eight different speakers were used in the analysis. Mobile phones now have a Punjabi ASR system thanks to Puneet and his colleagues. For feature extraction, the MFCC approach is used, and for pattern recognition, the GMM method is used.[86]. The 6.34 hour voice corpus data set with 48 distinct speakers was used. There are 1275 words in this speech corpus. For a higher degree of GMM, i.e., 64, the system provides the maximum accuracy. The GMM-CD United was the most accurate, with an accuracy of 81.2 percent. Ravindra and Ashok [4] presented and discussed various pattern recognition methods. As a result of this research, many classifiers used in the development of voice recognition systems have been evaluated. DBN was found to be superior to other classification methods after numerous algorithms were implemented and evaluated.

Table 4.1: State-of-the-art speech recognition models characteristics and challenges

Author [citation]	Methodology	Features	Challenges
Darekar and Dhande[62]	ANN	• Non-linear and complex relationships can be learned and represented, and the entire data set is saved on the network rather than the database.	• For building the structure, there is no rule; • it requires parallel-processing processors.
Patil and Lahudkar [63]	HMM	• Statistically sound; • can be incorporated into other databases.	• They can't specify dependencies between the concealed states, hence they tend to have more unstructured parameters.
Najnin and Banerjee [64]	Deep Neural Network	• A hierarchy of hidden layers captures non-linear relationships between the input variables, and the system is capable of doing multiple jobs at once.	• Performs poorly when dealing with little datasets.
Abdelaziz [65]	HMM	• Can represent a variety of Markov processes, including complex ones. • Utilize efficient algorithms for learning.	• Higher-order correlations are not captured. • The first-order Markov attribute restricts the number of first-order HMMs.
Tao and Busso [66]	Deep Neural Network	• Networks can uncover problems and provide results because they have the ability to learn on their own.	• It's not known how long the network takes to run.
Deng et al. [67]	Autoencoders	• Achieves peak performance • Improves the ability to recognize emotions even under challenging circumstances.	• The procedure takes a long time. • Training is really expensive.
Shi et al. [68]	SVM	• With unstructured data, it works well; with huge dimensional data, it scales rather well.	• The training of huge datasets takes longer because of this.

			• The final model is difficult to decipher and interpret.
Sharma et al. [69]	DNN-HMM	• Has a high level of efficiency. • HMM is capable of storing records with varying structures.	• In other words, HMM isn't fully automated. • Requires annotated data training.

4.2 Research Methodology

4.2.1 Comparing Different Techniques

This study looks at the five pattern recognition methods of KNN, SVM, NN, DNN, and DBN techniques. “To evaluate the performance of such approaches, we have used several measuring performance criteria, such as precision, sensitivity, FPR, FNR, FDR, MCC, etc., in comparing classification approaches, such as SVM, KNN, NN, and DBN”[99]. Compared with four other pattern identification processes, DBN works well with practically all pattern recognition criteria.

4.2.2 Optimization Techniques used for Deep Belief Network

Grey Wolf Optimizer

Mirjalili et. proposed the Grey Wolf Optimizer algorithm in 2016 [100]. In other words, it's based on the survival of the fittest. Evolution in nature is a result of processes like inheritance, selection, and mutation. Genetic inheritance, natural selection, and mutation are all used in the hunt for wolves. These can be found Wolve hunting is divided into four stages: I search for prey, II circle prey, III hunt, and IV attack on the target. Wolf fitness is sorted by ascending order, and then resolutions with fitness values lower than zero are removed, resulting in an equal number of newly produced wolves as the number of eliminated wolves.

Whale Optimization Algorithm

The Whale Optimization Algorithm was inspired by the Whale fish, the world's largest mammal[101]. The Whale Optimization Algorithm was developed using data from a humpback whale, one of seven kinds of whale. Bubble-net feeding is the basis of the

technique, which has only previously been observed in humpback whales. There are three steps to this strategy: encircling the prey, using a bubble-net attack, and hunting for prey in the last phase. WOA is a simple and reliable stochastic optimization method based on swarms. It is possible to avoid local optimum solutions by using population-based WOA.

Chaotic Biogeography Based Optimization

"The BBO algorithm imitates interactions in immigration, emigration, and mutation between distinct species(personalities) located in different habitats." [102]. In order to estimate the bio-geographical survival of this environment, the Environment Suitability Index (HBI), of each habitat is calculated. The high obstruction shows excellent adaptability for inhabitants and can accommodate numerous species. While the low score implies low habitat appropriateness, fewer species can survive in the habitat.

Rider Optimization Algorithm

The Rider Optimization Algorithm (ROA) is inspired by a group of Riders that desire to reach their destination and win[103][104]. Bypass riders, followers, overtakers, and attackers are the four sorts of riders. It's impossible to generalize about all the riders because they're all so different. Bypass riders use an alternate route to their destination, avoiding traffic on the main road. Instead of following the leading rider, it makes its own path. The bikers that follow the leader offer encouragement and support. The overtaker rides in the same direction as the leader until they reach their objective. Riders that attack using their verbal skills are known as attacker riders.[105]. The ROA algorithm is used to find the best weights in a neural network.

4.2.3 Proposed Architecture

The proposed system for developing the Marathi Speech Recognition System consists of a newly proposed hybrid feature extraction technique, the MFCC-SF, and the new hybrid classification approach RCBO-DBN. The pre-processing has been done using smoothing and filtering techniques to reduce the noise from the speech corpus. Figure 4.2 depicts the proposed system model.

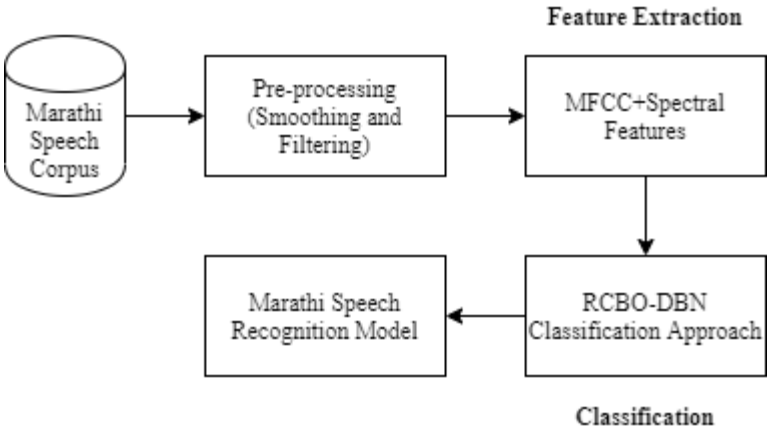


Fig. 4.2: Proposed Duo Features with Hybrid-Meta-heuristic-based Deep Belief Network

4.2.4 Duo-Feature Feature Extraction Technique

Speech recognition systems rely heavily on the feature extraction process. The accuracy and quality of features derived from speech samples as part of the SRS creation process impact the classifier's performance. 80% of the original speech corpus is used for training speech recognition system, while the remaining 20% is used as a test speech. MFCC[88], Spectral features[106], and Duo MFCC + Spectral features were used to extract the features. After performing this experiment, it is found that the Duo MFCC-SF hybrid approach is performing well compared to using MFCCC and Spectral features alone which are shown in Table 4.3. The extracted features are then passed through the DBN classifier, and the results are compared. Then, utilizing this feature extraction technique. 14 MFCC features are combined with 26 spectral characteristics in the proposed dual feature to help improve the accuracy and quality of features used in classification.

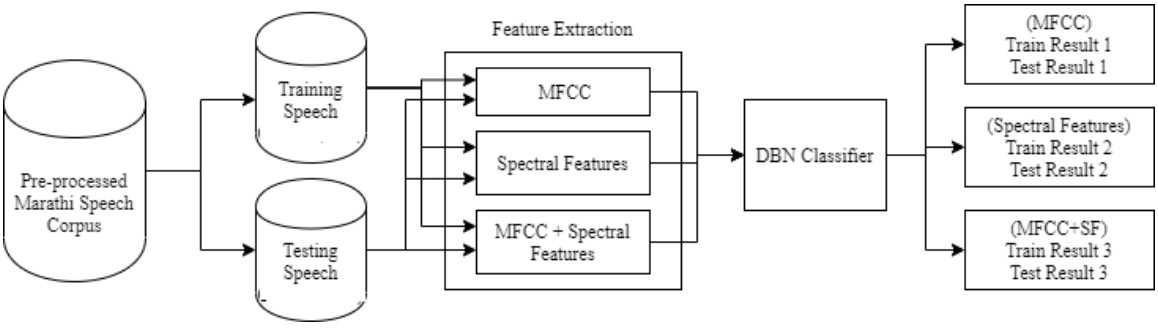


Fig. 4.3: Proposed Duo Features and conventional feature extraction approach

Algorithm 1: MFCC-SF Feature Extraction
Algorithm: MFCC-SF feature extraction

```
Input: Speech Signal
Output: Features
    function MFCC (speech signals)
        Initialize Parameters;
        Pre-emphasize the spectrum ;
        Apply hamming windowing to frames ;
        Get spectrum by applying FFT to all frames;
        Determine Metrix for mel-spaced filterbank;
        Transform spectrum to mel-specrum;
        Obtain MFCC vector for frame by applying discrete cuisine transform;
      end function
     function Spectral_Features (speech signals)
        input : speech signals;
        output : Spectral features
                Apply Fourier transform to convert input from time domain to
                frequency domain
                Extract spectral roll-off, spectral flux, and spectral centroid
   end function
   function MFCC_SF (speech signals)
        MFCC();
        PCA(MFCC features);
        Spectral_Features (speech signals);
   end function
end Algorithm
```

4.2.5 RCBO-DBN Pattern Recognition

The RCBO proposal combines Rider Optimization Method (ROA) and the Optimization-based Chaotic Biogeography (CBBO) algorithm. In one way or another, each of these approaches works well with the neural network. In order to get the most out of the hidden

neurons, the ROA method has been optimised for their optimal weights. The CBBO, on the other hand, is ideal for use with the Habitat Suitability Index (HBI). In order to create a Deep Belief Network's RCBO optimization algorithm, the algorithm's strength is evaluated and put to good use.

Algorithm 1:DBN Classifier
Input: Speech Signal
Output: Features
function DBN_Classifier (features)
Initialize Parameters;
Training the first RBM layer
Obtain the representation of input from first layer v or h_0
Input that representation to the next hidden layer h_1
Train the h1 layer as an RBM by taking input from previous RBM layer v or h_0
Repeat step 2 and step 3 for a defined number of layers.
End function
end Algorithm

Algorithm 2: Proposed Hybrid Optimizer RCBO
Input: Speech Signal
Output: no. of hidden neuron
function RCBO (features)
Initialize Parameters -pop, indiv, dim, num.gear;
Calculate Species Count
Calculate LamdaMu
Set steering position
Define Riders behavior
Compute convergrnce_curve
Initialize maximum rate for immigration and emigration
Compute best_fit

Return best_fit, Convergence_curve, leader_pos,ct

End function

4.3 Results and Discussion

4.3.1 Experimental Setup

The proposed system has been implemented in Python language. The speech corpus for the development of the Marathi SRS was collected from the Government of India's Indian Language Proliferation and Development Centre. This dataset contains approximately 44500.wav speech samples from around 1500 speakers having different gender and dialects. The neural network's hidden neurons are optimized to improve the DBN algorithms performance. The experiment was performed on a machine with 16 GB RAM and an Intel i7 processor. The speech corpus is divided into 6 equal parts as it is difficult to work with a single speech corpus on a machine with the above-mentioned configuration.

4.3.2 Performance Measure

The two metrics of type performance are evaluated, i.e., positive and negative. "Positive metrics include accuracy, sensitiveness, specificity, accuracy, predictive value (NPV), F1score, and correlation math coefficients (MCC). The negative measured are False Positive Rate (FPR), False-Negative Rate (FNR), and False Discovery Rate (FDR)"[99].

4.3.3 Performance Analysis of Various Approaches

The suggested system's performance analysis is divided into two parts (a) Performance evaluation of traditional and proposed feature extraction strategies. (b) Evaluation of various DBN optimization strategies in conjunction with the proposed RCBO-DBN classification methodology.

4.3.4 Performance Analysis of Feature Extraction Techniques

Table 4.2 shows the different parameters values of various approaches evaluated for the Marathi Language Speech Recognition System[4]. The purpose of this experiment is to understand how different approaches work if we use them for developing the Marathi SRS.

Based on the experiment, it was observed that Deep belief Network's overall performance was well.

Table 4.2 Different Approaches Comparison for Marathi SRS

Algorithms	Accuracy	Sensitivity	Specificity	Precision	FPR	FNR	NPV	FDR	F1 score	MCC
KNN	0.7395	0.9062	0.7361	0.0655	0.2638	0.0937	0.7361	0.9344	0.1222	0.2010
SVM	0.7086	0.9860	0.7029	0.0634	0.2970	0.0139	0.7029	0.9365	0.1192	0.2084
NN	0.7533	0.9300	0.7497	0.0705	0.2502	0.0699	0.7497	0.9294	0.1310	0.2159
DNN	0.7702	0.9678	0.7661	0.0778	0.2338	0.0321	0.7661	0.9221	0.1441	0.2378
DBN	'0.8148	'0.9356	'0.8123	'0.0923	'0.1876	'0.0643	'0.8123	'0.9076	'0.1681	'0.2605

4.3.5 Performance Analysis of Feature Extraction Techniques

The suggested MFCC-Spectral characteristics and standard MFCC and Spectral approaches for the extraction of feature are presented in Fig. 4.4. Fig. 4.4(a), MFCC extraction approaches, Spectral feature and proposed dual feature MFCC-Spectral are the first positive accuracy measures. The suggested dual feature provides a 12% improvement compared to MFCC and around 13% improvement compared to Spectral at an 85% learning level. Table 4.3 lists the Accuracy values for several datasets. The suggested Duo feature provides advances of 4 percent to MFCC and 5 percent to the spectral features displayed in Fig. 4.4(b) and the Precision values reported in Table 4.4. Compared to standard feature extraction methods; NPV is also a positive measure for performance analysis. Table 4.5 shows the NPV results of the proposed methodology for extraction of the feature (c). In relation to MFCC and Spectral characteristics, the findings are surprisingly poor. In comparison with standard approaches, the proposed strategy offers 10% underperformance.

The FPR is the first negative measure. Table 4.6 and Fig 4.4 shows values for FPR performance analyses for all methodologies (d). The FNR, the values referred to in

Table 4.7 and illustrated in Fig. 4.4, is another negative metric explored here (e). It provides improved performance of roughly 0.5% compared to MFCC and 1.5% compared to Spectral. For performance analysis the last negative measure is the FDR. Table 4.8 and Fig 4.4 exhibit the FDR findings of the proposed feature extraction approach (e). In relation to MFCC and Spectral characteristics, the findings are surprisingly poor. Compared to traditional approaches, the proposed method offers a 10% underperformance.

Table 4.3: Accuracy Performance Analysis

Algorithms	Testcase 1	Testcase 2	Testcase 3	Testcase 4	Testcase 5	Testcase 6	Mean of Result
MFCC	70.94	69.29	76.70	70.47	58.58	75.12	70.18
Spectral Features	69.06	70.79	76.54	72.21	61.11	72.03	70.29
Proposed Duo-Feature	83.12	82.66	86.00	81.24	74.29	84.68	81.99

Table 4.4: Precision Performance Analysis

Algorithms	Testcase 1	Testcase 2	Testcase 3	Testcase 4	Testcase 5	Testcase 6	Mean of Result
MFCC	6.35	5.90	7.43	6.20	4.54	7.02	6.24
Spectral Features	5.94	6.31	7.63	6.60	4.76	6.39	6.27
Proposed Duo-Feature	10.51	10.10	11.90	9.51	7.17	11.08	10.05

Table 4.5: NPV Performance Analysis

Algorithms	Testcase 1	Testcase 2	Testcase 3	Testcase 4	Testcase 5	Testcase 6	Mean of Result
MFCC	70.37	68.74	76.37	69.92	57.77	74.75	69.65
Spectral Features	68.48	70.23	76.13	71.68	60.38	71.56	69.74
Proposed Duo-Feature	82.79	82.36	85.84	80.89	73.78	84.47	81.68

Table 4.6: FPR Performance Analysis

Algorithms	Testcase 1	Testcase 2	Testcase 3	Testcase 4	Testcase 5	Testcase 6	Mean of Result
MFCC	29.62	31.26	23.63	30.08	42.23	25.25	30.35
Spectral Features	31.51	29.77	23.87	28.32	39.62	28.44	30.25
Proposed Duo-Feature	17.21	17.63	14.15	19.11	26.22	15.52	18.31

Table 4.7: FNR Performance Analysis

Algorithms	Testcase 1	Testcase 2	Testcase 3	Testcase 4	Testcase 5	Testcase 6	Mean of Result
MFCC	1.54	4.04	7.04	2.50	1.56	6.43	3.85
Spectral Features	2.43	1.73	3.39	2.00	3.01	4.89	2.91
Proposed Duo-Feature	0.97	2.89	0.75	1.58	0.67	5.24	2.02

Table 4.8: FDR Performance Analysis

Algorithms	Testcase 1	Testcase 2	Testcase 3	Testcase 4	Testcase 5	Testcase 6	Mean of Result
MFCC	93.65	94.10	92.57	93.80	95.46	92.97	93.76
Spectral Features	94.06	93.69	92.37	93.40	95.24	93.61	93.73
Proposed Duo-Feature	89.49	89.90	88.10	90.49	92.82	88.92	89.95

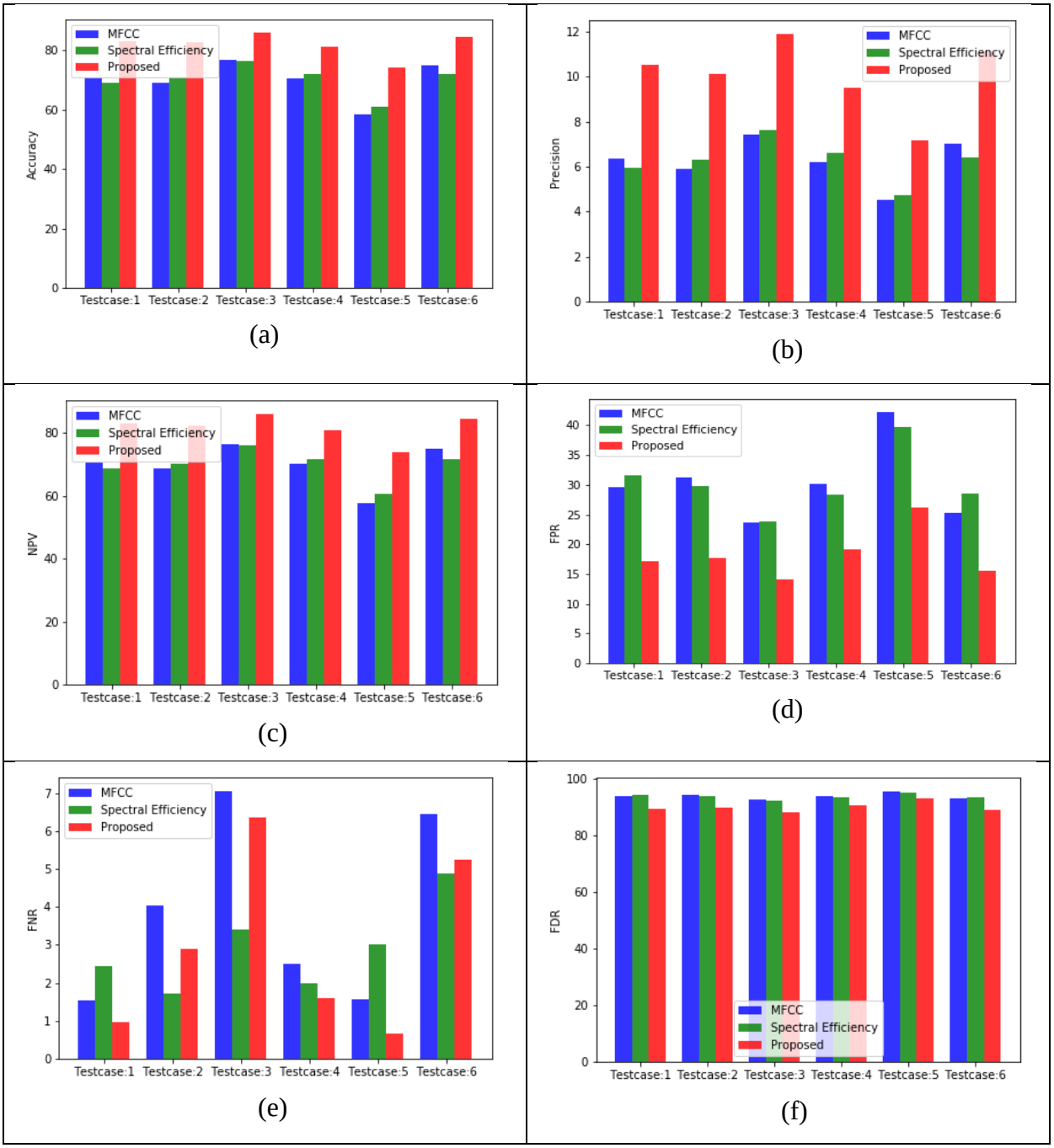


Fig 4.4. Performance analysis of various Feature Extraction Techniques

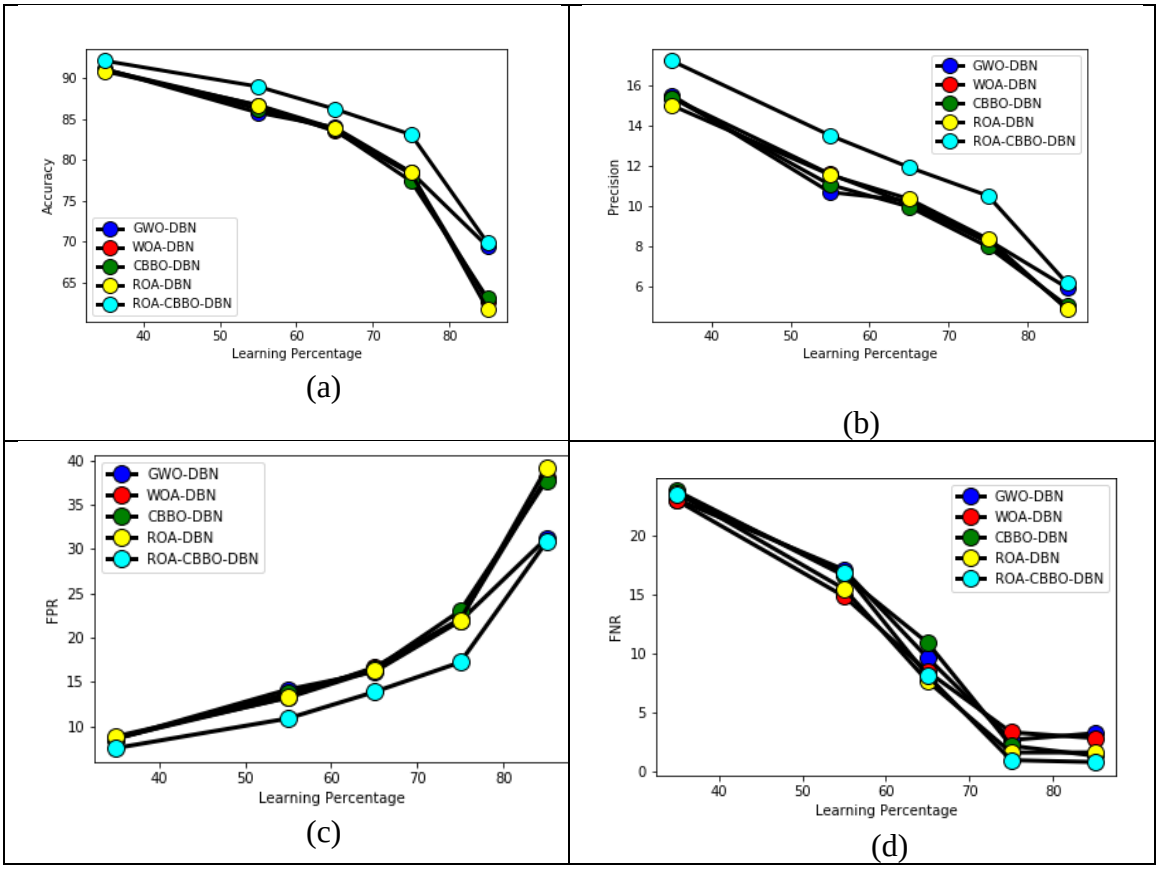
4.3.6 Performance Analysis of Conventional and Proposed RCBO Pattern Recognition Techniques.

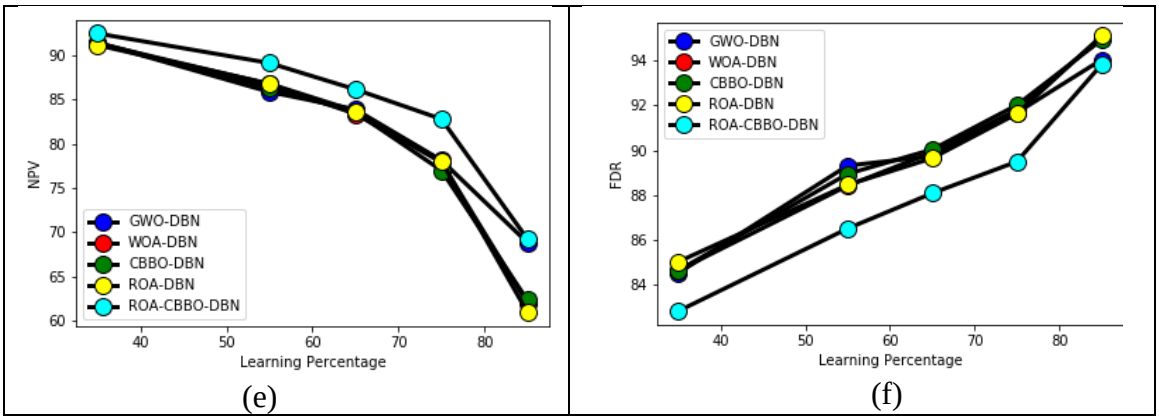
The performance of the proposed RCBO and the standard pattern recognition techniques GWO, WOA, CBBO, and ROA, is illustrated in Fig. 4.5, and the results are reported in Table 4.9. The initial positive measure accuracy of the suggested and standard optimization strategies is given in Fig. 4.5(a). The suggested approach has a greater accuracy of 71.5 percent than GWO-DBN, 8.84 percent than WOA-DBN, 8.46 percent than CBBO-DBN, and 9.82 percent than ROA-DBN. Table 7 displays these values. Precision values are graphed in Fig. 4.5(b). The precision values of the suggested optimized DBN provide a 0.22 percent improvement over GWO-DBN, 1.22 percent improvement over WOA-DBN, 1.1 percent improvement over CBBO-DBN, and 1.28 percent improvement over ROA-DBN. Another positive measure displayed in Fig. 4.5 is the Negative Predictive Value (c). The RCBO-DBN outperforms traditional optimised algorithms, advancing by 0.41 percent over GWO-DBN, 7.27 percent over WOA-DBN, 6.91 percent over CBBO-DBN, and 8.29 percent over ROA-DBN.

The first negative measure discussed here is the FPR, the values of which are shown in Table 4.9 and illustrated in Fig. 4.4 (d). It outperforms GWO-DBN by approximately 0.41 percent, WOA-DBN by 7.27 percent, CBBO-DBN by 6.91 percent, and ROA-DBN by 8.29 percent. FNR is another negative metric employed in this performance analysis. Table 4.9 and Fig. 4.5 exhibit the FNR findings of the suggested feature extraction approach (e). In comparison to other traditional optimized methods, the findings are promising. The suggested methodology outperforms GWO-DBN by 2.43 percent, WOA-DBN by 2.02 percent, CBBO-DBN by 0.54 percent, and ROA-DBN by 0.81 percent. FDR is the final negative measure employed in this performance study. Table 4.9 and Fig. 4.5 show the FDR results for the suggested and standard optimized approaches (f). The proposed technique outperforms GWO-DBN by 0.97 percent, WOA-DBN by 2.36 percent, CBBO-DBN by 0.54 percent, and ROA-DBN by 1.69 percent.

Table 4.9: Overall Performance Analysis of Optimized DBN Pattern Recognition Approach

Algorithms	Accuracy	Precision	NPV	FPR	FNR	FDR
GWO-DBN	69.39	5.95	68.82	31.17	3.24	93.65
WOA-DBN	62.66	4.95	61.96	38.03	2.83	95.04
CBBO-DBN	63.04	5.07	62.32	37.68	1.35	93.22
ROA-DBN	61.68	4.889	60.94	39.05	1.62	94.37
RCBO-DBN	71.50	6.17	69.23	30.76	0.81	92.68





(e) (f)

Fig 4.5: Performance analysis of various pattern recognition techniques

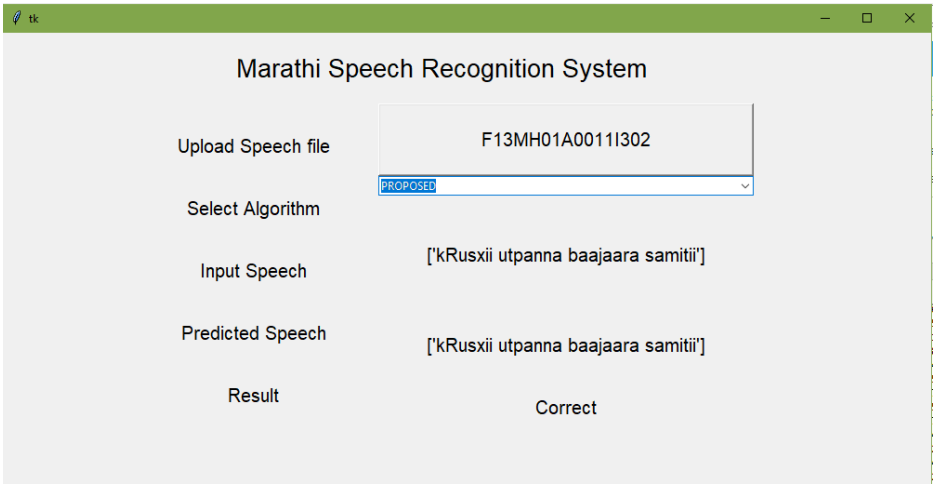


Fig 4.6 Marathi SRS Result Screen 1

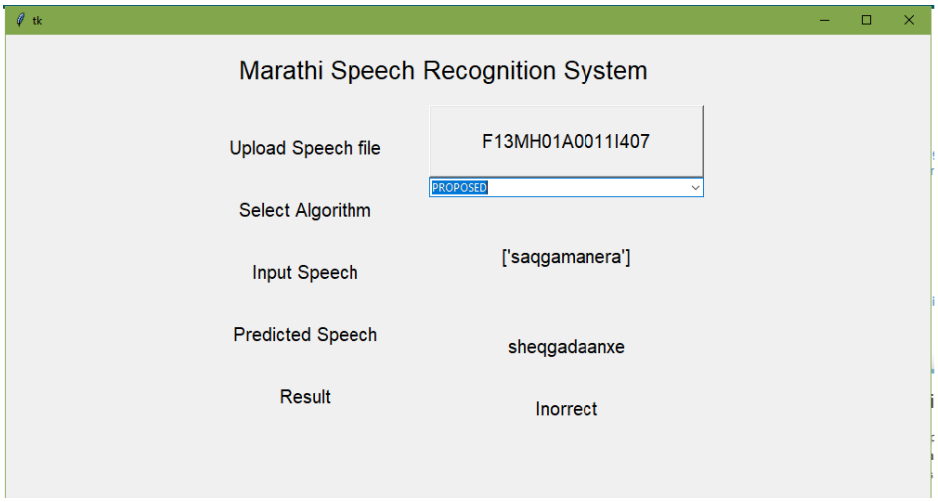


Fig 4.7 Marathi SRS Result Screen 2

4.4 Summary

During this study, we investigated several techniques for building a voice recognition system for Marathi. To evaluate the performances of these techniques using various metrics, including accuracy, sensitivity, FPR, FNR, FDR, and MCC, we examined classification methods, such as SVM, KNN, NN, and DBN. DBN performs well compared to all other techniques used for recognition. DBN gives about 81% accuracy. This study may be extended to increase the efficiency of buried neurons by changing the number and triggering the deep confidence network.

CHAPTER 5: ENHANCED MARATHI SPEECH RECOGNITION ENABLED BY GRASSHOPPER OPTIMIZATION-BASED RECURRENT NEURAL NETWORK

5.1 Introduction

In today's environment, speech recognition is a burgeoning and appealing research field[107]. Because of the difficulties of such languages, speech recognition in Indian regional languages is a widespread concern. Technological breakthroughs have emerged across the country without the bondage of language, which can be accomplished using the ASR model with regional languages[108]. The vocabulary that insists on operating the frequently used words often continues speech recognition. The term vocabulary refers to words that must be processed in linguistics. Marathi is an Indo-Aryan language spoken in central and western India that is phonetic in nature. However, there is a significant opportunity for constructing SRS models using Indian languages presented in various ways. The overall work in Indian regional languages has not yet been completed to a considerable degree by employing a true communication instrument like any other language. As a result, the research has been focused on the Marathi language. The primary goal of this study is to establish the Marathi ASR system's implementation.

Annotated transcriptions of speech are used to train an SRS model accurately. The amount of real annotated data that can be saved in reverberant and noisy situations is extremely limited, especially when compared to the amount of data that can be experimented with for cleaning annotated speech with the addition of noise. As a result, both simulated and actual data are employed extensively to improve robust speech recognition by increasing the diversity and quantity of training data. Furthermore, multi-task learning is used in the voice recognition model in reverberant and noisy environments. The training of an acoustic system is meant to solve two or more different tasks at the same time[109]. The auxiliary task's goal is to generate the features of clean speech using a regression loss.

The most efficient deep learning strategies for generating an SRS are based on supervised learning advances[16]. It is used to categorize chores. It is dependent on data annotation, which can be resource-intensive and time-consuming. Because of the acoustic environment's non-reverberant and clean character, the amount of annotated data available is crucial for SRS. However, these perfect acoustic settings are not as reasonable as in many real-life situations and may be harmed by deprivations of the speech signal caused by acoustic room proprieties or noise in the surrounding environment, which leads to reverberations of the speech. These occurrences can shape the speech recognition model, resulting in much lower outcomes and a considerably more complex process. Another issue in this reverberant and noisy instance is the limited amount of annotated real data.

This proposal is an essential contribution towards developing speech recognition system for the Marathi language.

- To create an RNN based speech recognition model utilizing GOA by various phases, such as preprocessing, extraction, and classification of features.
- The extraction of features using techniques like MFCC and spectral features such as spectral rod off, spectral centroid, and spectral flow.
- To optimize the numbers of hidden neurons in the RNN classification using GOA to design the Marathi SRS.

5.1.1 Related Work

Pironkov et al.[110] presented an HTL model in 2020. It frequently transitioned between single-task and multi-task learning by treating the input as either simulated or actual data. When using a denoising auto-encoder as auxiliary work in the configuration of a multi-task, hybrid design has enabled it to profit from both simulated and actual data. On the “CHiME4” database, this HTL design outperformed the traditional single-task learning strategy.

Sivaram and Nemala [111] used novel feedback and data-driven “discriminative spectro-temporal filters” to extract features in ASR in 2010. To begin, a set of "spectro-temporal filters" were devised to separate each phoneme from the remaining phonemes. For training the features derived from such filters, a hybrid HMM/MLP phoneme

identification model was used. The confusions were explicitly addressed for creating the second set of the filter in identifying the top confusions in the model. According to the results of the experiments, phoneme recognition established better features and contained more important matching information than standard voice recognition models.

Guglani and Mishra [112] recommended in 2020 that the performance of the ASR model be improved by incorporating the possibility of voicing computed features and pitch dependent characteristics. The pitch dependent features were used to perform the ASR model with Punjabi as the tonal language. As a result, the ASR model for the Punjabi language was constructed with the option of voicing computed features and pitch dependent characteristics. The ASR model's performance was tested using the WER measure, which considerably enhanced the characteristics. The WER was tested using features such as Kaldi pitch, FFV, SAcC, and Yin. The proposed Kaldi toolkit was utilized to outperform the other mentioned ASR models in terms of performance.

Huang and Renals [107] used a hierarchical Bayesian model to investigate an HPYLM in 2010. It has proposed a principled methodology for embedding the power-law distribution and smoothing a natural language model. Hierarchical Bayesian language models were used to experiment with conversational speech recognition, which resulted in significant decreases in WER and perplexity. As a result, the convergence of HPYLM is an important factor.

In 2021, Smit et al. used a variety of methodologies and technologies to achieve successful subword modelling for WFST-based hybrid DNN-HMM speech recognition via graphemes[113]. This model also assessed the four various languages, as well as the estimation of such approaches in a limited-resource setting. Furthermore, character-based NN-LM was used to investigate subword usage and concerns for hybrid DNN-HMM models. Finally, these techniques were evaluated across a variety of language modelling units.

5.2 Proposed Methodology

5.2.1 Proposed Architecture

The implementation of the Indian languages Speech Recognition model is difficult and hence allows for broad motivation and increasing study in this field. Most of the algorithms for deep education are still dependent on supervised learning in the field of SRS implementation. But supervised learning is based on annotating the data, a time-consuming and resource-intensive approach. A superior SRS model is created by a non-reverberant and clean auditory environment. It allows for a better model of language recognition. Although these ideal acoustic instances are not as realistic as they are in different real-life settings, the Speech Recognition model suffers from spoken signal degradations. Due to the character of each language, implementing the SRS model for many languages requires many methodologies. It has its own grammar and phonetic statements. The SRS model is a wider field of research for Indian regional languages such as Marathi. The key problem in building the speech recognition pattern is to reduce audio signal noise, which impacts the model's performance. Therefore, an efficient language recognition model for the Marathi language needs to be developed as indicated in the Fig. 5.1.

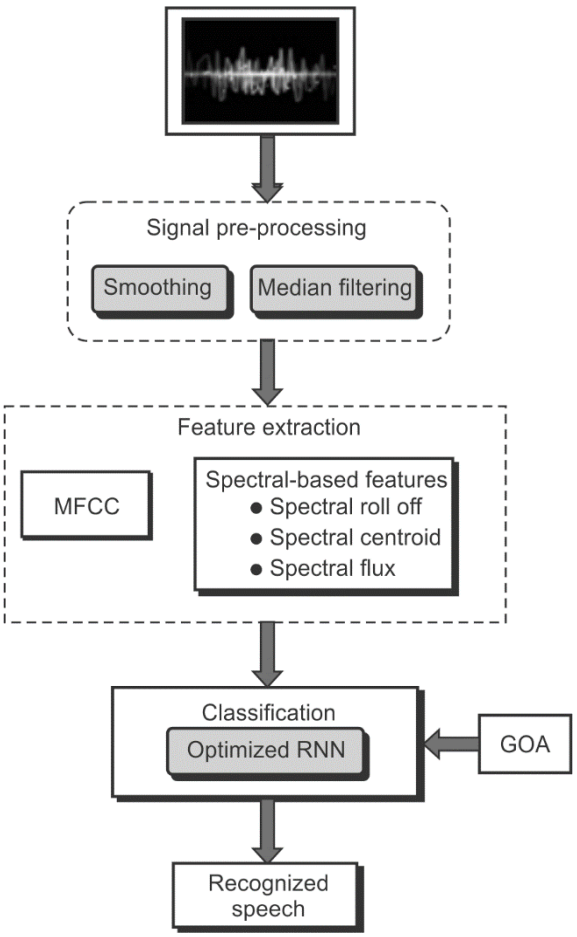


Fig. 5.1: Proposed GOA-RNN Speech recognition model on the Marathi language

The proposed Marathi language speech recognition model consists of three stages: preprocessing, feature extraction, and classification. The suggested speech recognition model receives the audio signal as input. The smoothing and median filtering procedures are used in the early step of preprocessing. Preprocessing is the process of removing artefacts and noise from a signal[114]. Second, the feature extraction procedure is used to extract MFCC and spectral-based features such as spectral rod off, spectral centroid, and spectral flux[99]. The feature extraction procedure is used to eliminate irrelevant information while retaining useful information. The extracted features are subjected to the classification step, which involves training with the extracted features from the input signal. The proposed speech recognition model on the Marathi language uses optimized RNN for feature classification, where the hidden neurons of the RNN are optimized using the GOA optimization algorithm[115]. The

proposed model achieved higher accuracy and precision in developing an efficient voice recognition model for the Marathi language.

5.2.2 Preprocessing of Input Speech Signals

This is the first stage of the suggested language recognition model in which the input is regarded as the signal. This signal is an analogue waveform which cannot be used with digital models immediately away. Preprocessing is therefore undertaken in order to turn the input talk into a form that is recognizable via the recognizer. The suggested model of voice recognition uses the two methods, smoothing and median filtering, outlined below.

Smoothing: It is a significant noise reduction method employed in the preprocessing stage of the proposed speech recognition model[114]. It reduces the audio signal noise, which improves the noise reduction performance. In smoothing, the signal data points are changed, so individual points are higher than nearby dots. The points below the next doors are maximized, resulting in a smoother signal.

Median filtering: It is a famous and well-known signal processing unit that is used mostly to denoise the input signal by removing certain noise. Equation (1) forms the median filtering.

$$MF = \begin{cases} y\left(\frac{D-1}{2}\right) & D\ odd \\ \frac{1}{2}\left[y\left(\frac{D}{2}\right) + y\left(\frac{D}{2}+1\right)\right] & D\,even \end{cases} \quad (1)$$

Here, the word MF refers to the average of a collection of two-part values. Consider a collection of D values sorted as y(d), where D is odd. The Easy element of MF middle value is considered as $y\left(\frac{D-1}{2}\right)$. Generally, median filters are selected for having the odd length as D.

5.2.3 Feature Extraction and Optimized RNN Adaptable for Proposed Speech Recognition Model

5.2.3.1 Feature Extraction Process

It is the second step of the proposed Marathi language speech recognition model for achieving various features. The main goal of feature extraction is to uncover a set of attributes of an utterance known as features. It is the process of obtaining important information while avoiding extraneous ones. This technique entails computing different critical signal characteristics such as frequency response or energy. Speech is defined as a succession of sounds with various qualities that are translated into smaller bits known as frames. As a result, these frames are processed in order to extract the significant features associated with voice vectors. The proposed model extracts MFCC and spectrum-based features as spectral rod off, spectral centroid, and spectral flux.

MFCC: MFCC is one of the main characteristics of the speech recognition model. It comes from the signal for the human input. The "linear cosine is a representation of a short-term log power spectrum of an anonymous mel frequency signal." MFCC extraction is of the kind that the full speech signal characteristics are concentrated in the initial few coefficients. The MFCC in Equation (2) is calculated.

$$Co_m = \delta_{Co} \sum_{f=0}^{F-1} \cos\left(m \frac{\pi}{F} \left(f + 0.5 \right) \right) \log_{10}\left(En_f \right) \tag{2}$$

The amplification factor is denoted by the word Co in Equation (2). Co_m denotes the dynamic range of coefficients as a function of the normalization factor, and En_f denotes the amount of energy in each channel in Equation (3).

$$En_f = \sum_{g=0}^{G-1} \sigma_f (g) X_g \tag{3}$$

The numbers of three-way filters utilized here 0 — pertaining to the F, G = 24 are counted as −f where |x(g)|2, and 0 — percentage of triangular filters are counted. The acquired MFCC features of the input signal are therefore further classified with the spectral functionality.

Spectral-based features: It is also known as frequency-based features since it is obtained by converting a time-based signal into a frequency domain using the Fourier transform. It is used to determine the pitch, notes, melody, and rhythm. Spectral roll-off, fundamental frequency, spectral flux, spectral density, frequency components, and spectrum centroid are some of the spectral-based properties. The suggested speech recognition model extracts the spectral centroid, spectral flux, and spectral roll off, which are discussed further below.

The spectral centroid is defined as the point at where the spectrum begins and ends. It is defined as "the power distribution centre of a signal's spectrum." It has various values for spoken and unspoken speech." It is the gravity of the spectrum, the sign function of which is given in Equation (4).

$$SC = \frac{\sum_{bn}^{NF/2} fr(bn) T_r[bn]}{\sum_{bn}^{NF/2} \left| T_r[bn] \right|} \tag{4}$$

The STFT frame of tr is showed as T_r [bn], the FFT point number is shown as NF, and the bin bn frequency is referred to as f_r (bn).

Roll-off spectral: It is "the skew of a frequency spectrum of the signal and the frequency below which a concentration of 85% of spectrum magnitude distribution is known as the roll-off" indicated in Equation (5).

$$\sum_{bn}^{NF/2} \left| T_r[bn] \right| \le 0.85 \sum_{bn}^{NF/2} \left| T_r[bn] \right| \tag{5}$$

Spectrum flux*:* It is defined as "a measure that characterizes the shift in the shape of a signal's spectrum." It is computed as the ordinary Euclidean norm of the delta spectrum magnitude" stated in Eq. (6).

$$SF_r = \sum_{bn}^{NF/2} \left(\left| T_r[bn] - T_{r-1}[bn] \right| \right)^2 \tag{6}$$

The proposed speech recognition model extracts the total number of features as 15.

5.2.3.2 Recurrent Neural Network

The 'end-to-end' structure for sequential data makes RNN an efficient way of recognising speech. RNN offers its perspective on the recognition of speech efficiently. However, RNN's success at recognising speech, as stated in the literature, is poor. This proposed model therefore uses optimized RNN for performance improvement. The

acclaimed optimization algorithm named GOA increases the RNN classification. RNN is meant to map or predict the sequence by sequence. Assume the sequence of input as () x =(x1), the hidden vector sequence is set to hs =(hs_1,...hs_{it}), and the output of the vector sequence is supplied as ot=(ot_1,....ot_{it})**.**

$$\vec{hs}_{it} = HS\left(wt_{xhs}x_{it} + wt_{hshs}hs_{it-1} + c_{hs}\right) \tag{7}$$

$$ot_{it} = wt_{hsot}hs_{it} + c_{ot} \tag{8}$$

The hidden layer function, or an element-wise application of a sigmoid function, is represented as HS, the weight matrices as wt, and the bias vectors as c. The logistic sigmoid function can be used to implement HS in this case.

$$ig_{it} = \alpha\left(wt_{xig}x_{it} + wt_{hsig}hs_{it-1} + wt_{vig}v_{it-1} + c_{ig}\right) \tag{9}$$

$$fg_{it} = \alpha\left(wt_{xfg}x_{it} + wt_{hsfg}hs_{it-1} + wt_{vfg}v_{it-1} + c_{fg}\right) \tag{10}$$

$$v_{it} = fg_{it}v_{it-1} + ig_{it}\tanh\left(wt_{xv}x_{it} + wt_{hsv}hs_{it-1} + c_{v}\right) \tag{11}$$

$$ot_{it} = \alpha\left(wt_{xop}x_{it} + wt_{hsop}hs_{it-1} + wt_{vop}v_{it} + c_{op}\right) \tag{12}$$

$$hs_{it} = ot_{it}\tanh\left(v_{it}\right) \tag{13}$$

In this section, the input, forget gate, gate, and cell vectors are represented respectively as ig, fg, op and v. BRNN has two various hidden layers that are transmitted to the same output layer. In BRNN, the front, the rear and the rear and the output are represented correspondingly as hs version, hs version and ot versus hs version. The output sequence is updated with € [IT,1] iteration for the iteration of the backward and forward layer of € [1,IT] in Eq. (14).

$$\vec{hs}_{it} = HS\left(wt_{x\vec{hs}}x_{it} + wt_{\vec{hs}\vec{hs}}\vec{hs}_{it-1} + c_{\vec{hs}}\right) \tag{14}$$

$$\overleftarrow{hs}_{it} = HS\left(wt_{x\overleftarrow{hs}}x_{it} + wt_{\overleftarrow{hs}\overleftarrow{hs}}\overleftarrow{hs}_{it+1} + c_{\overleftarrow{hs}}\right) \tag{15}$$

$$ot_{it} = wt_{\vec{hs}ot}\vec{hs}_{it} + wt_{\overleftarrow{hs}ot}\overleftarrow{hs}_{it} + c_{ot} \tag{16}$$

Use the deep-RNN idea created to create higher-level representations by stacking different hidden layers on one another and producing the input sequence for the next layer in the output sequence. For all L levels of a stack and hidden vector sequences Hs1 that is generated on a regular basis, the hidden layer function is utilized l € [1, L] and it €[1,IT].

$$hs_{it}^{l} = \left(wt_{hs^{l-1} hs^{l}} hs_{it}^{l-1} + wt_{hs^{l} hs^{l}} hs_{it-1}^{l} + c_{it}^{l} \right) \quad (17)$$

Here, assume $hs^0 = x$ and the network output ot_{it} is computed in Equation (18).

$$ot_{it} = wt_{hsLot} hs_{it}^{L} + c_{ot} \quad (18)$$

5.2.3.3 GOA for improved RNN

The proposed GOA-based RNN model uses efficient classification to optimise the number of hidden neurons in the RNN classifier with the GOA. This enhances the precise identification of the proposed model. GOA is used to resolve complications in real-time application structure optimization. It solves local optimum problems efficiently, gives acceptable results and explores search space. It also enhances optimal precision. It comes from the swarming tendency, both mature and nymph, of grasshopper insects. In the nymph phase, the grasshopper is actively engaged in little, sluggish stages while the adult phase activities are expanded and expedited. Equation simulates the swarming behaviour of the grasshopper. (19).

$$Gr_i = So_i + Gf_i + Wa_i \quad (19)$$

In Equation (19), the i^{th} grasshopper's position, social interaction, wind advection, and gravitational force on the i^{th} grasshopper are denoted as Gr_i,So_i,Wa_i, and Gf_i, respectively. In Equation (20), the swarming behaviour is adjusted by substituting the values of So_i and Gf_i.

$$Gr_i = \sum_{j=1, j\neq 1}^{N} so\left(\left|gr_j - gr_i\right|\right) \frac{gr_j - gr_i}{d_{ij}} - C\hat{e}_C + D\hat{e}_w \quad (20)$$

Here, the distance from i^{th} grasshopper and j^{th} grasshopper is called N, and the word so() is used to characterize the strength of the social pressures. $d_{ij}= |gr_j - gr_i|$ and the number of grasshoppers. $Gf_i = C\ddot{e}_c$, where ëc and C are referred respectively to as a unit vector to the centre of the earth and a gravitational constant. $Wa_i = D\ddot{e}_w$, where $\ddot{e}_w$ and D indicate respectively a unit vector in wind direction and a constant drift. Equation (20) has been amended to resolve the optimization concerns in Equation (21).

$$Gr_i^x = dc\left(\sum_{j=1, j\neq 1}^{N} dc \frac{up_x - lo_x}{2} s\left(\left|gr_j^x - gr_i^x\right|\right) \frac{gr_j - gr_i}{gr_{ij}} \right) + \hat{T}A_x \quad (21)$$

A decreasing coefficient is given as dc in Equation (21), which is defined in Equation (22), TA_x is the target value in the X^{th} dimension, and the upper and lower bounds in the X^{th} dimension are represented as up_x and lo_x, respectively.

$$dc = dc\max - it\frac{dc\max - dc\min}{IT} \tag{22}$$

Equation (22) indicates the maximum value and minimum value of dc, dcmax and dcmin, and the maximum number of iterations and current iteration, respectively, are recorded in IT. dcmax and dcmin values are assumed as both 1 and 0.00001. The GOA algorithm pseudo-code is shown in algorithm 1.

Algorithm 1: GOA [16]

Initialization of grasshopper swarm $Gr_i(i = 1,2,3,\ldots\ldots,n)$
Initialization of variables
Compute the fitness of every search solution agent
Consider the best search solution agent as TA
While $(it < IT)$
  Update dc by Eq. (22)
  for each search solution agent
    Normalization of distances between the grasshoppers in [1,4]
    Update the location of the current search agent solution using Eq. (21)
  End for
  Update TA if there is the best search solution agent
  $it = it + 1$
End while
Return TA

Objective function: The proposed Marathi language speech recognition model employs optimized RNN-based on GOA to maximize accuracy and precision for effective recognition. As shown in Equation 23, the proposed speech recognition model's fitness function should optimize accuracy and precision.

$$ff = \arg\max_{\{HN\}}\left(ACC + \Pr()\right) \tag{23}$$

In this context, the term ff indicates the model's fitness function. Here is the word HN number in the classification of RNN hidden neurons. The solution is from 5 to 55. The

accuracy is shown as ACC and accuracy is referred to as Pr. Precision ACC is the "ratio of observation of all observations precisely anticipated," and Precision Pr is a "ratio of positive observation predicted accurately with the overall number of positive observations" indicated in Equation (24) and Equation (25) respectively.

$$ACC = \frac{(pot + pon)}{(pot + pon + fap + fan)} \tag{24}$$

$$\Pr = \frac{pot}{pot + pot} \tag{25}$$

True positives, true negatives, false positives, and false negatives are denoted by the abbreviations pot, pon, pot, fap, and fan, respectively.

5.3 Results and Discussions

5.3.1 Experimental Setup

Python was used to implement the proposed model of speech recognition for the Marathi language. The proposed model considered 25 iterations and 10 populations as the maximum number. In comparison with the GOA-RNN algorithm, the performance of a proposed model was assessed using several methods such as RNN. LSTM and CNN. The dataset collected from the Linguistic department of the Government of India is divided into six equal speech corpus as it is difficult to work on the data set with mentioned hardware. The experiment was performed on the machine having 16 GB RAM and an Intel i7 processor.

5.3.2 Performance Metrics

Various performance measures are used for evaluating the proposed speech recognition model using the optimized RNN with GOA, which is described below. The performance metrics used for analyzing the performances of the algorithms are Accuracy, Sensitivity, Specificity, FPR, FDR, FNR, and WER.

5.3.3 Performance Analysis

Table 5.1. Overall Performance Analysis for Speech corpus 1

Measures	Precision	Accuracy	Specificity	FNR	Sensitivity	MCC
RNN [116]	0.088714	0.879724	0.883557	0.883557	0.342584	0.657416

LSTM [93]	0.128746	0.915134	0.918934	0.918934	0.305263	0.694737
Res-CNN [117]	0.150365	0.917875	0.919446	0.919446	0.173206	0.826794
RCBO-DBN [99]	0.105114	0.831192	0.827946	0.827946	0.009732	0.990268
GOA-RNN	0.135838	0.925823	0.93093	0.93093	0.370335	0.629665

Table 5.2. Overall Performance Analysis for Speech corpus 2

Measures	Accuracy	FPR	Specificity	FNR	Sensitivity	MCC
RNN [116]	0.877079	0.118337	0.881663	0.356164	0.643836	0.215921
LSTM [93]	0.913878	0.082195	0.917805	0.285959	0.714041	0.297147
Res-CNN [117]	0.917079	0.080866	0.919134	0.1875	0.8125	0.343111
RCBO-DBN [99]	0.826601	0.176349	0.823651	0.028878	0.971122	0.282361
GOA-RNN	0.921073	0.072789	0.927211	0.391267	0.608733	0.266914

Table 5.3. Overall Performance Analysis for Speech corpus 3

Measures	Accuracy	FPR	Specificity	FNR	Sensitivity	MCC
RNN [116]	0.903293	0.093039	0.906961	0.282407	0.717593	0.280643
LSTM [93]	0.932377	0.0647	0.9353	0.215608	0.784392	0.368508
Res-CNN [117]	0.93599	0.06244	0.93756	0.143519	0.856481	0.408533
RCBO-DBN [99]	0.860038	0.141524	0.858476	0.063421	0.936579	0.305622
GOA-RNN	0.941717	0.053895	0.946105	0.280423	0.719577	0.367503

Table 5.4. Overall Performance Analysis for Speech corpus 4

Measures	Accuracy	FPR	Specificity	FNR	Sensitivity	MCC
RNN [116]	0.876544	0.118839	0.881161	0.359756	0.640244	0.213388
LSTM [93]	0.912504	0.08381	0.91619	0.276132	0.723868	0.297783
Res-CNN [117]	0.916628	0.081767	0.918233	0.165505	0.834495	0.350011
RCBO-DBN [99]	0.812421	0.191084	0.808916	0.01586	0.98414	0.274065
GOA-RNN	0.921436	0.072763	0.927237	0.375436	0.624564	0.273722

Table 5.5. Overall Performance Analysis for Speech corpus 5

Measures	Accuracy	FPR	Specificity	FNR	Sensitivity	MCC
RNN [116]	0.849565	0.144644	0.855356	0.439773	0.560227	0.1602
LSTM [93]	0.88029	0.114237	0.885763	0.393182	0.606818	0.207357
Res-CNN [117]	0.888829	0.109188	0.890812	0.210227	0.789773	0.287862
RCBO-DBN [99]	0.742921	0.262189	0.737811	0.006689	0.993311	0.22877

GOA-RNN	0.897592	0.094292	0.905708	0.507955	0.492045	0.182202

Table 5.6. Overall Performance Analysis for Speech corpus 6

Measures	Accuracy	FPR	Specificity	FNR	Sensitivity	MCC
RNN [116]	0.897497	0.098897	0.901103	0.280875	0.719125	0.27498
LSTM [93]	0.925734	0.070759	0.929241	0.247706	0.752294	0.34195
Res-CNN [117]	0.931916	0.066493	0.933507	0.146789	0.853211	0.399461
RCBO-DBN [99]	0.846811	0.155245	0.844755	0.052448	0.947552	0.294547
GOA-RNN	0.934629	0.060243	0.939757	0.318984	0.681016	0.333555

The suggested speech recognition model is evaluated on the basis of a study of accuracy, sensitivity, specificity, FPR, and FNR with six speech corpora, as shown in Fig. 5.2. The results of the analysis are presented in Table 5.1 to Table 5.6. As previously stated in the experimental setting, both the traditional and proposed algorithms are tested with the entire six-speech corpus in question. This section contains the evaluation results, which are organized into Tables 5.1 through 5.6, with a graphical depiction of the data provided in Fig. 5.2. The results in Fig. 5.2 and in Tables 5.1 to 5.8 are on a scale from 0 to 1, and the percentages are explained in the context of the performance discussion.

The proposed voice recognition model is evaluated using an exactness analysis of several test scenarios, as illustrated in Fig. 5.2. GOA- RNN's accuracy, for test case 1, with learning percentage 85, is 5.4 percent, 4%, and 1.7% better than RNN, LSTM, and Res-CNN; the accuracy of the proposed model in test cases 2, 8 percent, 5.4 percent, and 3%, respectively, are improved on RNN, LSTM, and Res-CNN by the learning percentage of 85. In test case 3, the GOA-RNN outperforms RNN, LSTM, and Res-CNN by 9.3%, 1.7%, and 0.5%, respectively, and is 85 percent more accurate. The accuracy of the proposed GOA-RNN model for test case 4 is 6%, 3.6 percent, and 1.7 percent greater than the accuracy of the GOA-RNN model for the RNN, LSTM, and Res-CNN percentages of 85. In test case 5, GOA-RNN accuracy is 2.4 percent greater than RNN, 3.6 percent higher than LSTM, and 1.1 percent higher than Res-CNN in terms of learning percentage 85. At a learning level of 85, the suggested model outperforms RNN, LSTM, and Res-CNN by 9 percent, 2.9 percent,

and 1.7 percent, respectively. As a result, when compared to existing methodologies, the proposed speech recognition model enhances precision.

As a result, when compared to previous methods, the accuracy of the suggested speech recognition model is improved significantly. In terms of the parameters used to evaluate performance, such as accuracy, sensitivity, specificity, FPR, and FNR, the suggested optimized RNN method employing ROA outperformed the RNN, LSTM, and CNN in comparison to the other algorithms. However, none of the techniques for parameter MCC performs satisfactorily in terms of performance. In addition, the Python-based speech recognition system was constructed, and the user can check the many models that were employed here through the use of a graphical user interface.

As a result, when the overall parameters results are taken into account, the proposed model outperforms the others.

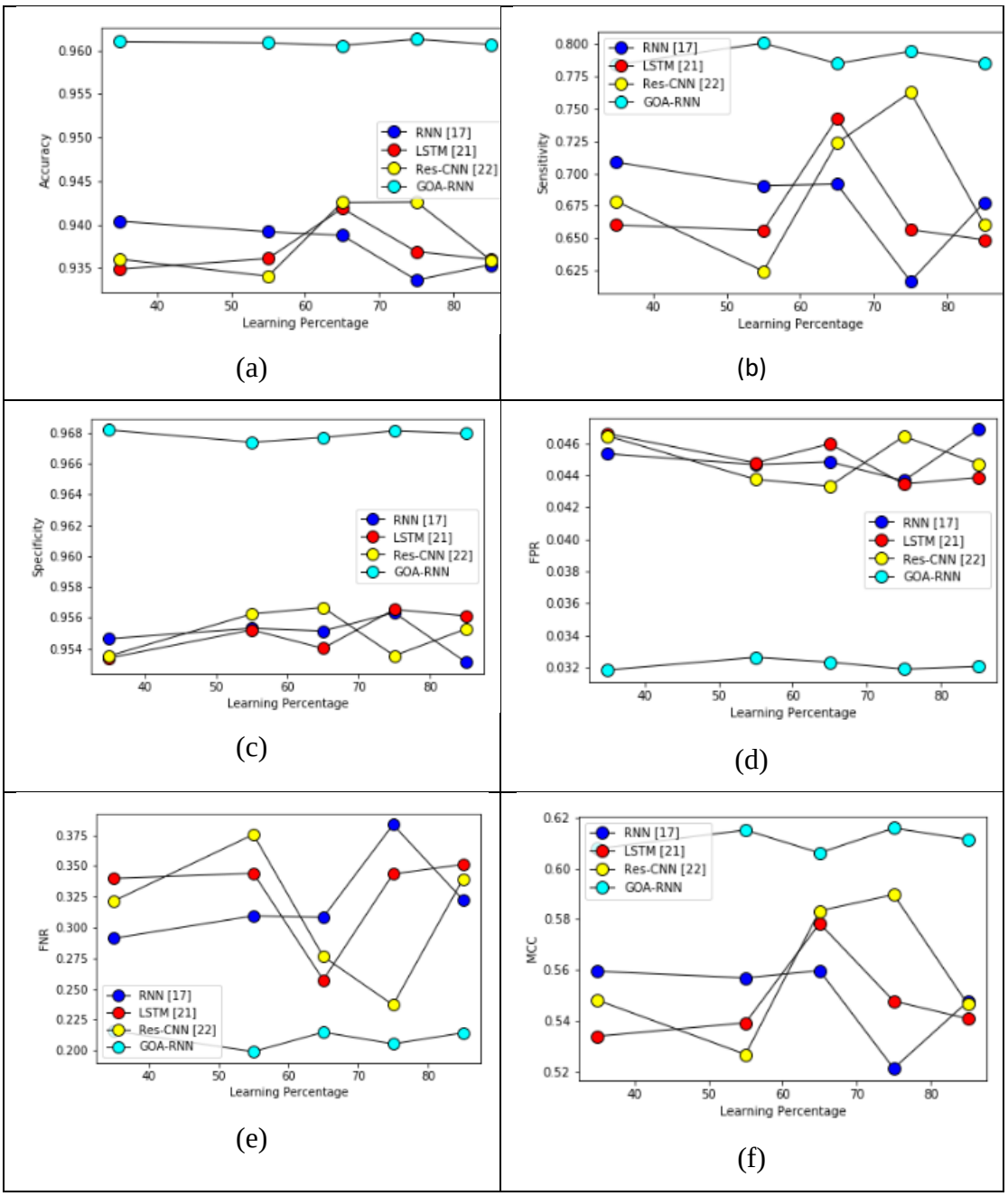


*Fig. 5.2. Performance analytical model for (a) Speech Corpus 1 (b) Speech Corpus 2, (c) Speech Corpus 3, (d) Speech Corpus 4, (e) Speech Corpus 5 and (f) Speech Corpus.*6

Table 5.7. Analysis of WER for the Proposed Speech Recognition Model for Diverse Speech corpus

Measures	Speech corpus 1	Speech corpus 2	Speech corpus 3	Speech corpus 4	Speech corpus 5	Speech corpus 6
RNN	12.0276	12.2921	9.6707	12.3456	15.0435	10.2503
LSTM	8.4866	8.6122	6.7623	8.7496	11.971	7.4266
CNN	8.2125	8.2921	6.401	8.3372	11.1171	6.8084
GOA-RNN	7.4177	7.8927	5.8283	7.8564	10.2408	6.5371

The study is conducted out on the WER measure in order to demonstrate the efficiency of the suggested speech recognition model, which is shown in Table 5.7 for the entire six-speech corpus tested. The GOA-RNN outperforms the RNN, LSTM, and CNN in terms of accuracy by 4.6 percent, 1.06 percent, and 0.79 percent, respectively. RNN, LSTM, and CNN are all significantly inferior to the GOA-RNN on the second speech corpus, with the GOA-RNN outperforming them all by 4.39 percent, 0.71 percent, and 0.39 percent, respectively. GOA-RNN outperforms RNN, LSTM, and CNN on the third speech corpus, improving by 3.84 percent, 0.93 percent, and 0.57 percent over their respective baselines. For speech corpus 4, the GOA-RNN outperforms RNN, LSTM, and CNN by 4.48 percent, 0.89 percent, and 0.48 percent, respectively, and outperforms RNN by 0.89 percent. For the speech corpus 5, the GOA-RNN outperforms RNN, LSTM, and CNN by 4.80 percent, 1.73 percent, and 0.87 percent, respectively, with RNN, LSTM, and CNN being the least improved. For speech corpus 6, the GOA-RNN outperforms RNN, LSTM, and CNN by 3.71 percent, 0.88 percent, and 0.57 percent, respectively, and outperforms RNN by 0.88 percent. Furthermore, the WER for the suggested model is examined for the remainder of the Speech corpus in a similar manner. As a result, the suggested speech recognition model incorporating GOA-RNN has been shown to produce superior outcomes when measured using the WER metric.

5.4 Summary

In this chapter, we proposed a new speech recognition model for the Marathi language that is based on RNN-based GOA and uses RNN-based GOA. Three stages were involved in the development of this model, including preprocessing, feature extraction, and classification. The input signals were preprocessed before being exposed to the feature extraction stage, which was the final step. MFCC and spectral-based features were extracted for use in the proposed speech recognition model, and the results were presented here. These characteristics were identified using an optimized RNN, where the number of hidden neurons was tuned using GOA to reduce the number of false positives. Finally, the proposed model has successfully efficiently achieved the recognized speech. Consequently, based on the experimental results, the WER of the proposed model was 3.84 percent, 1.06 percent, and 0.79 percent higher than that of RNN, LSTM, and CNN for speech corpus 1, and it achieved results similar to those of the remaining speech corpus.

CHAPTER 6: BIG BANG BIG CRUNCH BASED LSTM APPROACH FOR DEVELOPING MARATHI SPEECH RECOGNITION SYSTEM

6.1 Introduction

Because of the flexibility and convenience provided, several hands-free speech communication models have been used in a variety of applications, such as automatic speech recognition and multi-microphone portable devices in recent years [113]. While the examined speech signals are modified by room reverberation, background noises, and interfering speakers, the automatic identification model's performance is often minimized [118]. Furthermore, many speech enhancement approaches aim to reduce noise without changing speech signals to improve the robustness and performance of recognition models [119]. On the other hand, speech recognition models are challenging due to several constraints such as freestyle or spontaneous speech, a lack of dependability to speech differentiations such as speaking rate, gender, sociolinguistics, accents, and environmental noise. To overcome the issues in these models, it is necessary to bridge the gap between voice recognition methods and humans.

Due to the complexities involved in the classification of languages with a shared origin and the intermixing of distinct languages and multilingual datasets, speech recognition models are critical. As a result, there is a need to address existing recognition systems' limits to get optimal outcomes. As a result, certain study studies have taken into account one of the Indian language, such as Marathi. There is, however, little proof for providing helpful solutions when recognizing the Marathi language. On the other hand, the researchers were inspired to focus on constructing a new intelligent recognition model due to the lack of an efficient Marathi speech identification model and its local significance. Furthermore, the Marathi language model suffers from a lack of appropriate datasets and tiny vocabulary systems.

For speech recognition fields, many deep learning algorithms perform well. These approaches are utilized for automatic recognition models in "single-channel speech

enhancement," which improves recognition performance. Various speaker adaption approaches are being developed in present studies by targeting diverse speakers. While present deep learning algorithms frequently provide additional benefits, they also suffer from computational and language difficulties. Furthermore, the scarcity of research on Marathi language models motivates academics to focus on constructing a new Marathi language voice recognition model.

The main contribution of the proposed model is presented below.

- To demonstrate a new Marathi language speech recognition model including several steps such as pre-processing, feature extraction, and classification utilizing a heuristic-based classification technique.
- To improve performance, extract valuable features from speech signals using MFCC and spectrum-based features such as spectral rod off, spectral centroid, and spectral flux. Using the Principal Component Analysis (PCA) approach, the obtained features are reduced to get important features.
- To maximize recognition accuracy, optimize the number of hidden neurons and weight in the LSTM classifier using the PB3C method for obtaining recognized speech signals.

6.1.1 Related Work

A novel approach for creating sub-word language systems was proposed by Smit and colleagues [113] in 2021, which takes into account "Deep Neural Networks (DNN), Hidden Markov Models (HMM)," and weighted finite-state transducers." It has been proposed in this research to use an acoustic system that includes character models and sub-word language systems without the need for pronunciation dictionaries. A number of methods have also been developed to combine the advantages of different types of language model units, such as reconstruction and combination recognition lattices, in order to improve overall performance. The Neural Network Language Models (NNLMs) were produced using the developed model, which was viable due to the limited number of input and output layers.

The four languages, including "Finnish, Swedish, Arabic, and English," were used to evaluate distinct sub-word units on the Speech corpus, which included "Finnish, Swedish,

Arabic, and English." The experimental analysis was carried out, and the results showed that the results were more consistent and that the error rate had been lowered.

To obtain complete information, Hui et al. [118] created a novel "Iterative Mask Estimation (IME)" structure for upgrading the "complex Gaussian mixture model (CGMM)-based beamforming" technique in 2019. This model has built a "Neural-Network (NN)-based ideal ratio mask estimator" that has been trained from a multi-condition Speech corpus in order to incorporate the prior information into the current model. Following that, voice activity prediction information was obtained from the speech recognition results in order to make use of the rich context information in language models and deep acoustic, which was then used to reduce the insertion errors and refine the mask estimate, among other things. As part of the testing phase, the built model was subjected to the "CHIME-4 Challenge ASR problem of identifying 6-channel microphone array speech" which was created by Chime-4. The experimental findings have demonstrated that the proposed IME approach consistently and considerably outperforms the existing CGMM method while also lowering the error rate.

In 2017, Kipyatkova and Karpov[119] developed a recurrent artificial neural network-based automatic voice recognition model for the Russian language that was successfully implemented. Within the hidden layer, it has taken into account a diverse number of elements, and the baseline trigram language model has been created using linear interpolation of neural network models. It was determined how well the generated model performed based on its WER rating.

A new "DNN-based acoustic modelling" framework for the ASR model was introduced in 2015 by Zhou et al.[120], which used multiple DNNs (mDNN) to compute the posterior probability of HMM states and was implemented by Zhou et. al. Initially, the HMM states were classified into separate disjoint clusters based on the data-driven methodologies that were taken into consideration.

Then, the mDNN was trained to cluster the states into several categories. According to the researchers, they demonstrated that the contemplated training approach employing the mDNN model was utilized to boost the training speed, which included sequence-level discriminative training and frame-level cross-entropy training. The performance of the developed model has been improved as a result of the suggested model.

A DNN-based ASR model was developed by Xue et al. [121], who presented different layers of "pre-trained DNN" employing a novel group of connection weights to create a novel ASR model in 2014. Furthermore, from the adaption data, the training approaches have acquired a new condition code for each and every test condition that has been encountered.. The quick adaption technique employed in this model's development of an ASR model with supervised speaker adaptation was derived from a previous model. Also, by comparing the suggested adaptation scheme with various methodologies, they were able to conduct an experimental examination of the proposed adaptation scheme. Last but not least, they have achieved higher performance in terms of WER, accuracy, and precision, among other things.

6.2 Proposed Methodology

6.2.1 Proposed Model and Description

In recent years, there has been an increase in the number of research papers focusing on speech recognition models utilizing machine learning methodologies. Deep learning techniques are being applied in a variety of speech-related applications. Due to the complexity in determining local languages and the association between multiple languages, speech recognition models are critical concerns. As a result, deep learning-based approaches to speech recognition in the Marathi language are required, as shown in Fig. 6.1.

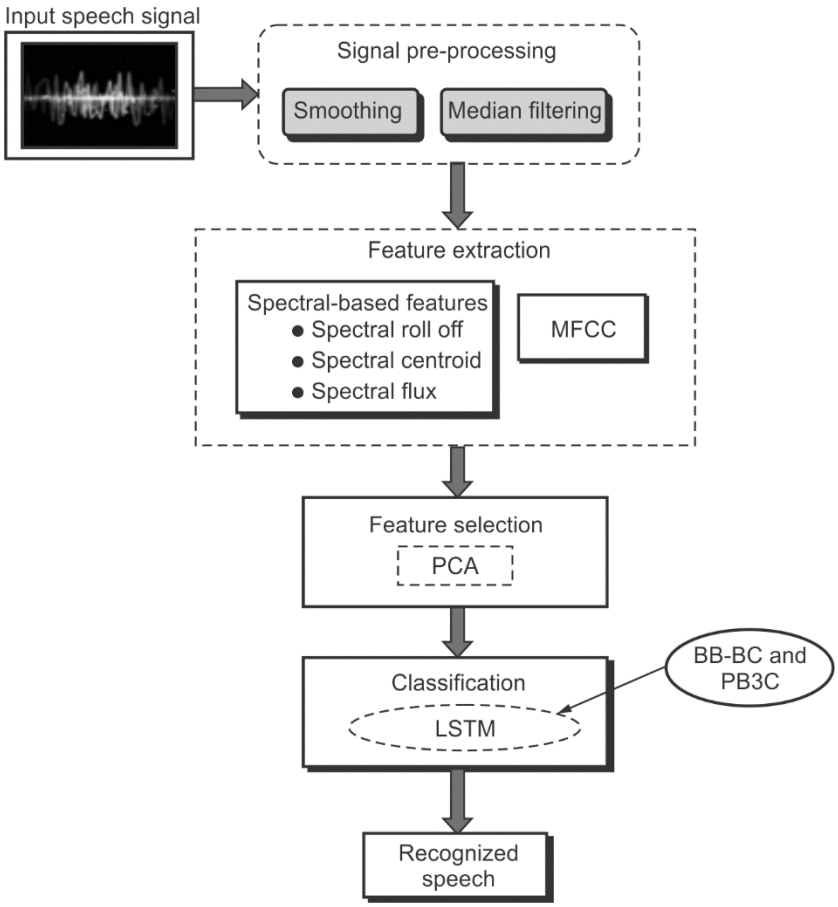


Fig. 6.1: Architectural representation of the proposed speech recognition model

The stages involved in the proposed Marathi language speech recognition model are "pre-processing, extraction, selection of the characteristics and classification." The audio signals collected are transmitted to the preprocessing stage using smoothing and median filtering techniques. It is intended to remove artefacts and sounds from the signals obtained. The second stage is feature extraction. Here the MFCC-SF feature extraction methodology proposed by Ravindra et al.[99] In addition, the feature reduction has been implemented on obtained features using PCA to reduce the sized data collected. The features picked are sent to the classification/ pattern recognition step, in which the LSTM algorithm with PB3C is combined[122] and checked the performance with pure LSTM and BB-BC-LSTM[123]. Finally, it uses the PB3C-based LSTM approach to achieve recognizable voice signals. Use the P3BC optimization technique to optimize the hidden neurons of LSTM. In order to be an effective language recognition model for Marathi language, this model seeks to improve classification accuracy.

6.2.2 PB3C-LSTM Approach

It is necessary to provide the PB3C-LSTM model with the PCA-selected features in order for it to recognize speech signals in the Marathi language efficiently. The PB3C technique is used to optimize the number of hidden neurons in the LSTM in this example. The accuracy of speech-recognized signals is intended to be improved by using this model.

In general, LSTM [124] is referred to as "variation in a recurrent network employing memory blocks," and it is a type of neural network. The LSTM is composed of three layers: the "input layer," the "hidden layer," and the "output layer." It is also known as memory blocks with information, and it is these cells that are used in the hidden layer where the Adam training model is applied. The LSTM is composed of three gates, which are labeled as "input gate, output gate, and forget gate." In a block of memory cells, the terms "input gate" and "output gate" are used to govern the input and output functions of the cells. A forget gate is then added after that, where the LSTM network is used to obtain the unit activations from a sequence of inputs, where and signify the number of features from PCA, and the output is used to find a mapping between the two sets of activations obtained.

$$i_n = \alpha\left(wg_{im} Fs_n^{PCA} + wg_{iq} q_{n-1} + w_{ih} h_{n-1} + v_i\right) \quad (1)$$

$$f_n = \alpha\left(wg_{fm} Fs_n^{PCA} + wg_{fq} q_{n-1} + wg_{fh} h_{n-1} + v_f\right) \quad (2)$$

$$h_n = f_n \otimes h_{n-1} + i_n \otimes g\left(wg_{hm} Fs_n^{PCA} + wg_{hq} q_{n-1} + v_h\right) \quad (3)$$

$$p_n = \alpha\left(wg_{pm} Fs_n^{PCA} + wg_{pq} q_{n-1} + wg_{ph} h_{n-1} + v_p\right) \quad (4)$$

$$q_n = p_n \otimes jh_n \quad (5)$$

$$o_n = \varphi\left(wg_{pq} q_n + v_n\right) \quad (6)$$

Specifically, the forget gate bias vector and the output gate bias vector are denoted as v_f and v_h, respectively, the input vector or current time step is denoted as Fs_n^{PCA}, the input gate bias vector and the input gate are both denoted as and, and the input gate is denoted as and, respectively, in the aforementioned equations. Furthermore, the forget gate and the weight matrices are denoted as fg and wg, respectively, and the cell activation vector, the output gate, and the previous output from the blocks are denoted as p, h and $p_{(n-1)}$, respectively.

The cell activation vector, the output gate, and the previous output from the blocks are denoted as, j , go and α , respectively, for the network output activation function, the cell output functions, and the cell input functions. It is denoted as φ for the network output activation function and for the network output activation function. In addition, the "hyperbolic tangent (tanh) activation functions" are used in the multilayer LSTM framework to enhance its performance. Additionally, the words and represent, respectively, the output of the current blocks and the memory of the current blocks, as well as shown in the Table 6.1, the diagonal weights of peepholes connections are supplied as terms, and, as shown in this table, the "highest weight of the input gate to the input" is given as, and reflects the output from the preceding memory from input blocks. The PB3C algorithm is used to optimize the number of hidden neurons in the new LSTM, which results in the use of the new LSTM. Its goal is to increase the classification accuracy to obtain correct voice signals in the Marathi language, among other things.

This method is prompted by the "Big Bang Theory" in cosmological research, which explains the origin of the universe as an explosion. The algorithm is based on this theory. During the huge crunch stage, the population is randomly distributed based on the location of the centre of mass. According to the BB-BC technique, the initial candidates are uniformly distributed throughout the search space. The big bang phase is followed by the big crunch stage, in which the fitness functions of each candidate and the current positions are replaced by the convergence operator, resulting in the production of a weighted average point, which is referred to as a centre of mass, as defined in Equation (7).

$$c_{a,b}^{y} = \frac{\sum_{z=1}^{np} \frac{1}{fh_b^z} c_{a,b}^{z}}{\sum_{z=1}^{np} \frac{1}{fh_b^z}} \tag{7}$$

Equation (7) denotes the fitness function of the solution at the iteration, the component of the solution at the iteration is denoted as, and the component of the centre of a mass point at the iteration is denoted as, the total number of candidates in the population is denoted as, the current iteration is denoted as, and the current dimension is denoted as it is believed that the most recent centre of mass is deemed to be the core in the following iteration, which then explodes during the big bang phase. The explosion also results in the production

of new candidates since the usual distribution of mass is followed exactly about the centre of mass, as stated in this formulation. (8)

$$c_{a,(b+1)} = c_{a,b}^{z} + \frac{rn_a \times \delta \times (c_{\max} - c_{\min})}{(b+1)} \quad (8)$$

When a random integer from the standard normal distribution is used in Equation (8), the upper and lower limits are denoted as $c_{\max}$ and $c_{\min}$, respectively,. In addition, a parameter δ is used to limit the parameters of the search space, and the new candidate is denoted as $c_{a,(b+1)}$. Fixing the value of standard deviation from Equation (8) produces the best results, and the standard deviation is fixed for the purpose of inversely lowering the current iteration produces the best results. This phase of the big crunch contraction is utilised for recalculating the "centre of mass," which is done after the big bang explosion has completed its cycle. Until the termination requirement is reached, the processes of explosion and contraction are continuously repeated in an endless loop. This BB-BC algorithm seeks to get the best possible results in speech recognition in the Marathi language, with the primary goal being the maximisation of recognition accuracy as the primary goal. In Algorithm 1, the pseudo-code for the BB-BC algorithm is described in detail.

Algorithm 1: BB-BC algorithm [123]
Initialization of random number, population, and iteration
Compute the fitness of every search agent
Estimate center of mass by Equation (7)
While $(b < B)$
Create new solutions by Equation (8) around the center of mass
Compute the fitness of every search agent
Update new solutions
End while
Terminate

PB3C algorithm [18]: This is an enhanced version of the BB-BC algorithm, which is a multi-population optimization algorithm that performs significantly better than the BB-BC algorithm in terms of accuracy and convergence rate. In order for this process to succeed, the elite must be updated by taking into account the population's local best answer. The solutions are revised with the help of Equation (9).

$$c^{\to y} = \frac{\sum_{z=1}^{np} \frac{1}{fh_b^z} c^{\to z}}{\sum_{z=1}^{np} \frac{1}{fh^z}} \quad (9)$$

The center of mass is used to determine which individual is the most physically fit. Furthermore, "the new candidate solutions are updated around the center of mass as a result of removing or adding a normal random number" that is decreased as the number of iterations elapses, as specified in Equation (1). (10).

$$c_{new(z,k)} = \zeta_{best(z,k)} + \frac{(c_{\max} * rn)}{l} \quad (10)$$

In this case, the maximum number of iterations is denoted by the symbol, and a random number is denoted by the symbol. PB3C pseudo-code is illustrated in Algorithm 2, which is a representation of the algorithm.

Algorithm 2: PB3C algorithm [122]

Initialization of' N' population and each population consist of 'C' candidate solutions

Compute the fitness of every candidate

Set $i = 1$

While (i < TC)

for $b = 1:N$

for j=1:C

Compute the fitness of j^{th} candidate solution.

end for

Calculate the local best of b^{th} population

end for

Amongst the 'N' local best solutions, find out the global best

With the given probability, move the local best candidate solution towards the global best

for $b = 1:N$

Create new population around local best candidate solution

end for

i =i+1

End while

Terminate

Objective Model (C)

The accuracy of the proposed speech recognition model for the Marathi language, which is based on the P3BC-LSTM, is maximised in order to provide exact recognition. According to Equation (1), the suggested speech recognition model's objective function is formulated as follows: (11).

$$fh = \underset{\{HN\}}{\arg\max}(Ac) \tag{11}$$

The word "fitness function" refers to the fitness function of the proposed speech recognition model in the Marathi language, which is defined as follows: Specifically, the word refers to the "number of hidden neurons in the LSTM," which can range between 5 and 55 in this context. Precision is represented as, which is defined as a "ratio of exactly anticipated observations to the total number of observations" as provided in Equation (12).

$$Ac = \frac{\left(po^{true} + po^{neg}\right)}{\left(po^{true} + po^{neg} + fa^{true} + fa^{neg}\right)} \tag{12}$$

The terms "true positives," "true negatives," "false positives," and "false negatives" are used to refer to the po^{true}, po^{neg}, fa^{true} and fa^{neg}, respectively.

6.3 Results and Discussions

6.3.1 Experimental Setup

Python was used to implement the proposed speech recognition model for Marathi. The experiment is performed on a machine having 16 GB RAM and an Intel i7 processor. The suggested model considers the maximum number of iterations to be 25 and the population to be 10. On the six speech corpora, the performance of the proposed PB3C-LSTM model was evaluated using the LSTM and BB-BC. The no. of hidden neurons and weights are optimized for better results.

6.3.2 Performance Metrics

For the evaluation of performance, the following are described: The many performance measures.

(a) **WER:** It is used for measuring the proposed speech recognition model.

$$WER(\%) = \frac{De + Ns + Ie}{Nw} * 100(\%) \quad (13)$$

Here, the terms NS , DT , IE and NT denotes "the number of substitutions in test, the number of deletions in the test, the number of insertion error in the test and the number of words utilized in a test", respectively.

(b) **Word Accuracy Rate (WAR):** It is derived in Equation (14).

$$WAR(\%) = \frac{Nw - Ds - Ns}{Nw} \quad (14)$$

(c) **Sentence Error Rate (SER):** It is correlated among audio predicted correctly to total number of audio as given in Equation (15).

$$SER(\%) = 1 - \frac{PA}{TA} \quad (15)$$

Here, terms PA and TA represents the audio signals predicted correctly and the total number of audios, respectively.

6.3.3 Performance Analysis

The performance of the proposed PB3C-LSTM approach and other approaches such as GOA-RNN, LSTM, and BBBC-LSTM is mentioned in Table I and shown in Fig. 6.2, Fig. 6.3, and Fig. 6.4.

Table 6.1: Performance analysis of the proposed speech recognition model on the Marathi language with different algorithms for six different datasets in terms of error measures

Algorithms	WER	WAR	SER
Speech Corpus 1			
LSTM [93]	0.253333	0.826667	0.233934
ROA-RNN [125]	0.22	0.833333	0.220825
BB-BC-LSTM [123]	0.233333	0.846667	0.221979
PB3C-LSTM[122]	0.193333	0.86	0.189498
Speech Corpus 2			
LSTM [93]	0.206667	0.86	0.190093
ROA-RNN [125]	0.253333	0.833333	0.2414
BB-BC-LSTM [123]	0.22	0.853333	0.206988

PB3C-LSTM[122]	0.206667	0.853333	0.231292
Speech Corpus 3			
LSTM [93]	0.266667	0.826667	0.26818
ROA-RNN [125]	0.226667	0.833333	0.20578
BB-BC-LSTM [123]	0.213333	0.866667	0.23345
PB3C-LSTM[122]	0.24	0.84	0.254593
Speech Corpus 4			
LSTM [93]	0.226667	0.853333	0.225522
ROA-RNN [125]	0.206667	0.866667	0.213203
BB-BC-LSTM [123]	0.226667	0.853333	0.230098
PB3C-LSTM[122]	0.22	0.853333	0.200145
Speech Corpus 5			
LSTM [93]	0.22	0.833333	0.223583
ROA-RNN [125]	0.266667	0.82	0.289717
BB-BC-LSTM [123]	0.233333	0.826667	0.23908
PB3C-LSTM[122]	0.226667	0.866667	0.230612
Speech Corpus 6			
LSTM [93]	0.22	0.866667	0.215973
ROA-RNN [125]	0.2	0.866667	0.223778
BB-BC-LSTM [123]	0.2	0.866667	0.216662
PB3C-LSTM[122]	0.18	0.88	0.194015

Analysis on SER by varying learning percentage

The proposed Marathi language speech recognition model utilising P3BC-LSTM is evaluated in terms of SER by altering the learning percentages for six datasets, as shown in Fig. 6.2. On dataset 1, the SER of the developed P3BC-LSTM is 44 percent, 39.5 percent, and 34.7 percent higher than that of the LSTM, ROA-RNN, and BB-BC-LSTM, which are all 35 percent.

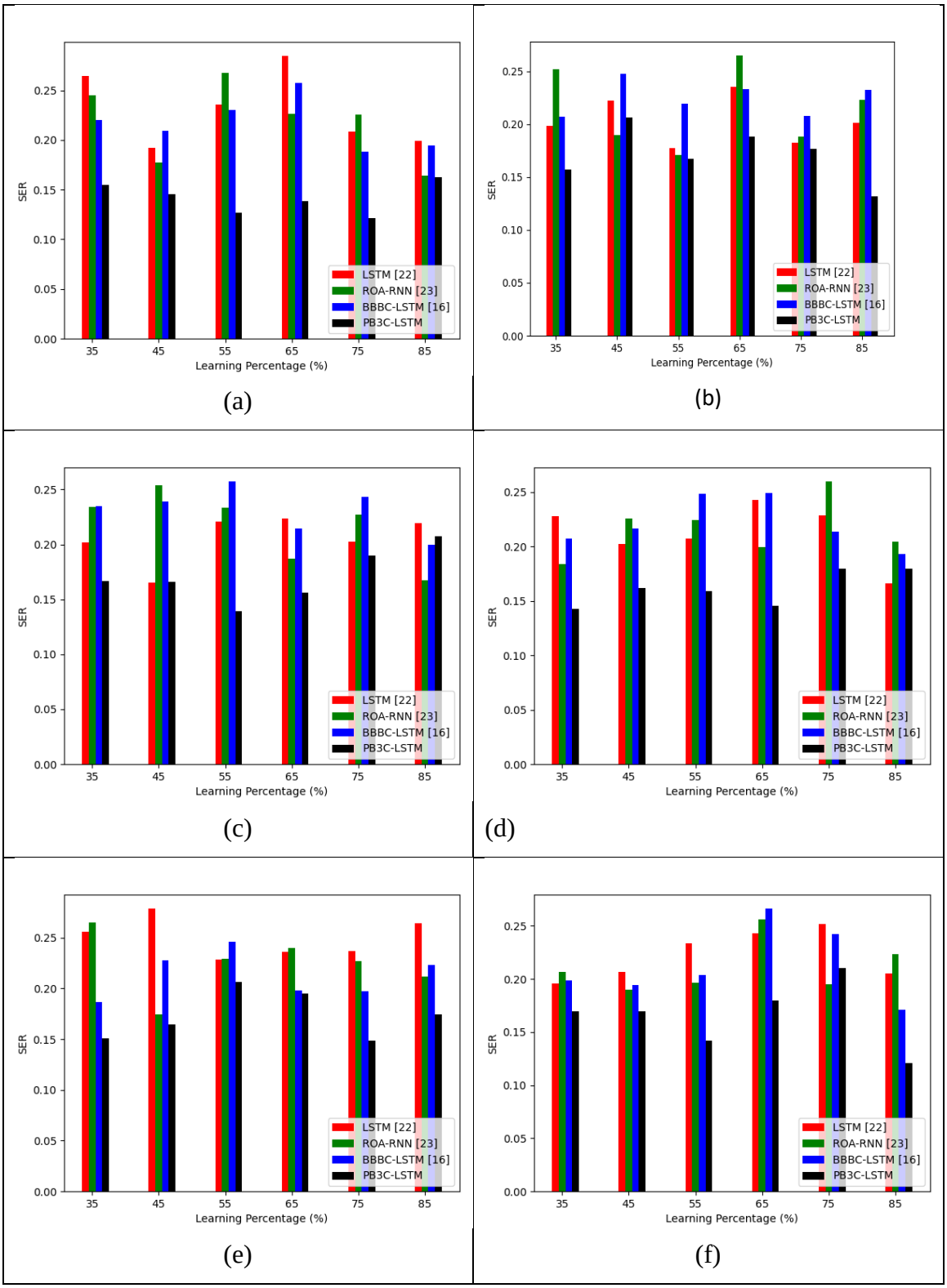
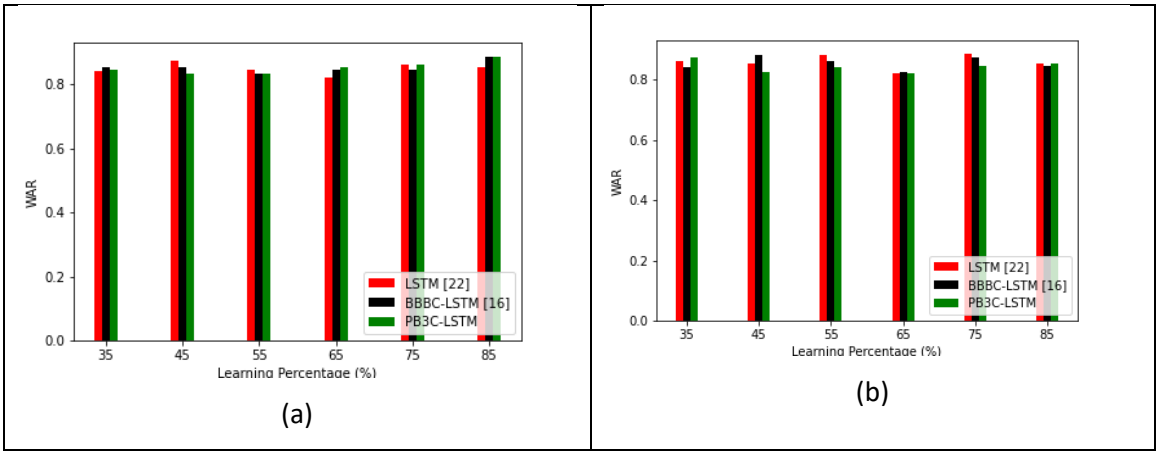


Fig 6.2: Performance analysis of the proposed speech recognition model on the Marathi language in terms of SER with different conventional approaches for "(a) Dataset 1, (b) Dataset 2, (c) Dataset 3, (d) Dataset 4, (e) Dataset 5 and (f) Dataset 6"

On dataset 2, the performance of the suggested P3BC-LSTM is 11% better than LSTM, 5.8 percent better than ROA-RNN, and 30.4 percent better than BB-BC-LSTM at 55%. The generated P3BC-LSTM has an SER of 27 percent, 5.8 percent, and 23.8 percent higher than the LSTM, ROA-RNN, and BB-BC-LSTM, which are all at 65 percent for dataset 3. On dataset 4, the performance of the designed BB-LSTM is 27 percent better than LSTM, 30.4 percent better than ROA-RNN, and 36 percent better than BB-BC-LSTM, with a score of 55 percent. Similarly, the SER of the proposed P3BC-LSTM on dataset 5 is 37.5 percent, 31.8 percent, and 25 percent higher than that of the LSTM, ROA-RNN, and BB-BC-LSTM, which are all 75 percent. Finally, on dataset 6, the accuracy of the considered P3BC-LSTM is 38 percent, 43.4 percent, and 23.5 percent greater than LSTM, ROA-RNN, and BB-BC-LSTM, which are all at 85 percent. As a result, the proposed speech recognition model outperforms the existing approaches in terms of SER for the six datasets.

WAR Analysis by varying learning percentage

The performance of the generated WAR model is examined as shown in Fig. 6.3. The WAR of the suggested PB3C-LSTM has increased by 7.4 percent, 5.8 percent, and 8.4 percent compared with LSTM, ROA-RNN and BB-BC-LSTM. The projected WAR of PB3C-LSTM is 2.2% by 4.0%, and 6.9% better by 55% compared with LSTM, ROA-RNN and BB-BC-LSTM by 2.5% correspondingly. WAR is 9.5%, 12%, and 12.2% higher than LSTM, ROA-RNN, and BB-BC-LSTM correspondingly, taking into account dataset 3 with a value of 65%.



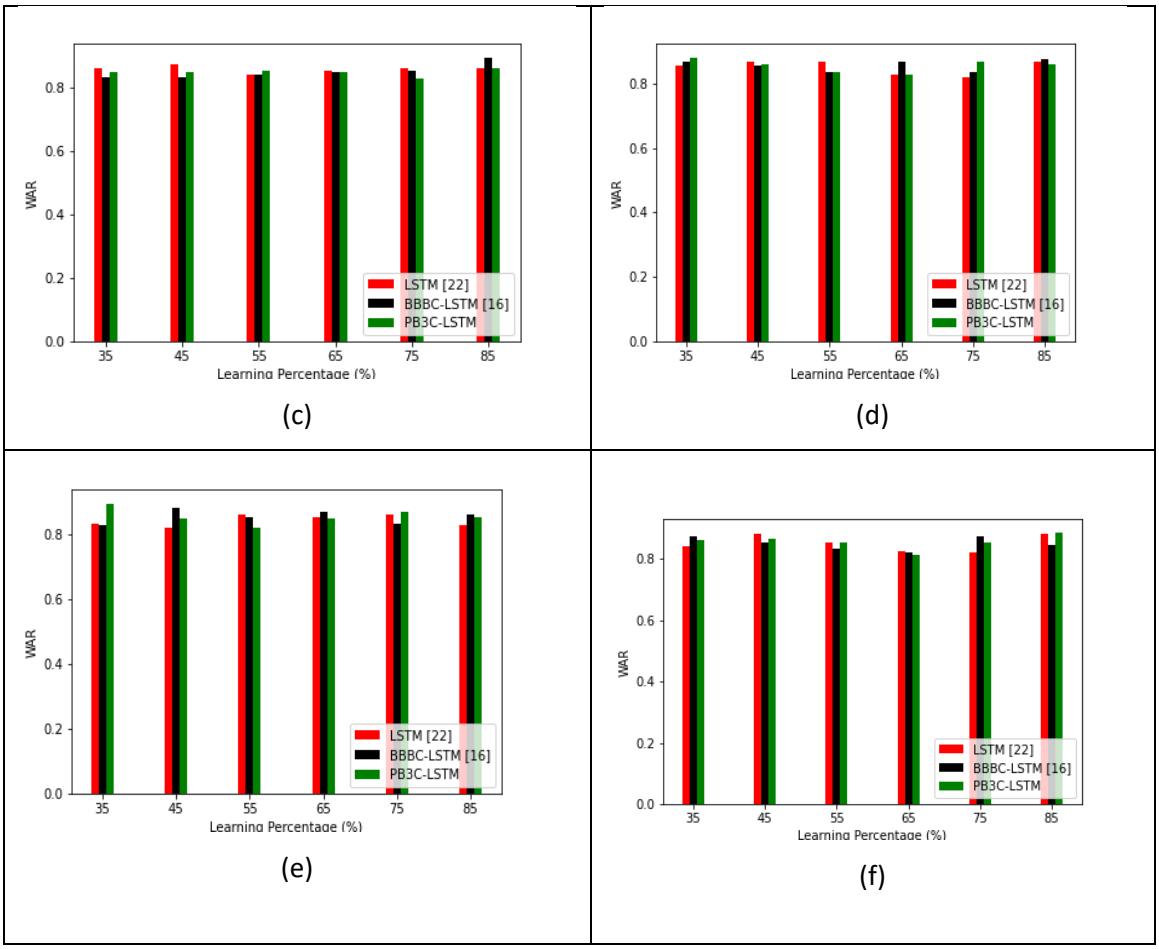
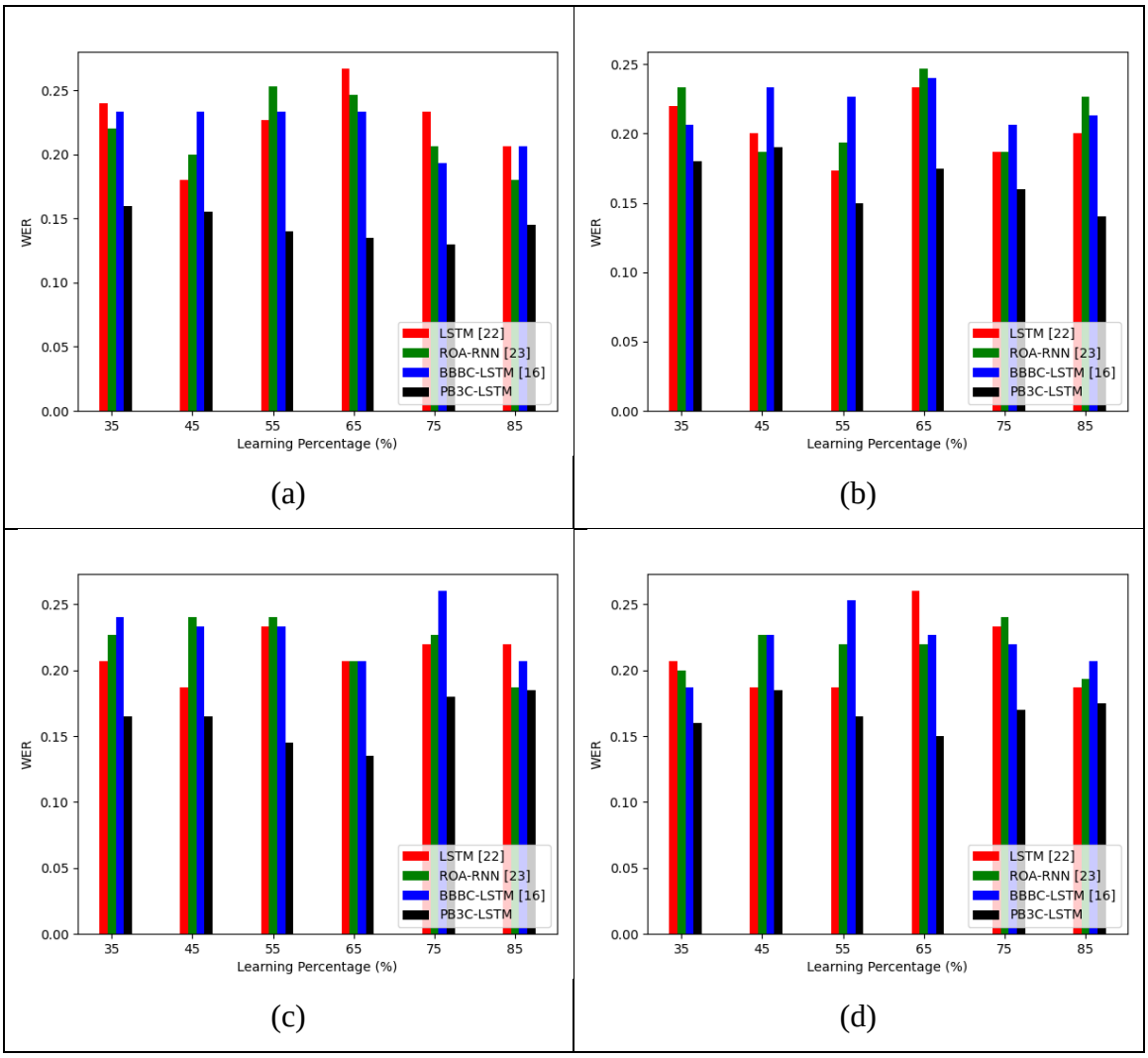


Fig 6.3: Performance analysis of the proposed speech recognition model on the Marathi language in terms of WAR with different conventional approaches for "(a) Dataset 1, (b) Dataset 2, (c) Dataset 3, (d) Dataset 4, (e) Dataset 5 and (f) Dataset 6"

For dataset 4, the WAR performance of the implemented PB3C-LS is 9.5%, 6.9%, and 4.5% better than LSTM, ROA-RNN, and BB-BC-LSTM, and 65%, respectively. With data set 5 being considered at 75%, the WAR for the PB3C-LSTM applied is 9.5%, 15%, and 8.2% higher than the LSTM, ROA-RNN, and BB-BC-LSTM, respectively. Also, the performance of the PB3C-LSTM developed is 9.5%, 15%, and 9.5% better than LSTM, ROA-RNN and BB-BC-LSTM, respectively, assuming WAR data set 6 at 85%. Thus the performance of the created P3BC-LSTM voice recognition model in WAR compared with other techniques is validated.

WER Analysis by varying learning percentage

The constructed speech recognition model's performance is evaluated with WER by altering the learning percentages, as shown in Fig 6.4. While examining dataset 1, the WER of the recommended PB3C-LSTM is 34.6 percent, 30.4 percent, and 33.3 percent higher than LSTM, ROA-RNN, and BB-BC-LSTM, respectively. While examining dataset 2, the WER of the examined PB3C-LSTM is 14 percent, 21%, and 37.5 percent better than LSTM, ROA-RNN, and BB-BC-LSTM, respectively. While examining dataset 3, the WER of the used PB3C-LSTM is 5.8 percent, 34.6 percent, and 33.3 percent better than LSTM, ROA-RNN, and BB-BC-LSTM, respectively.



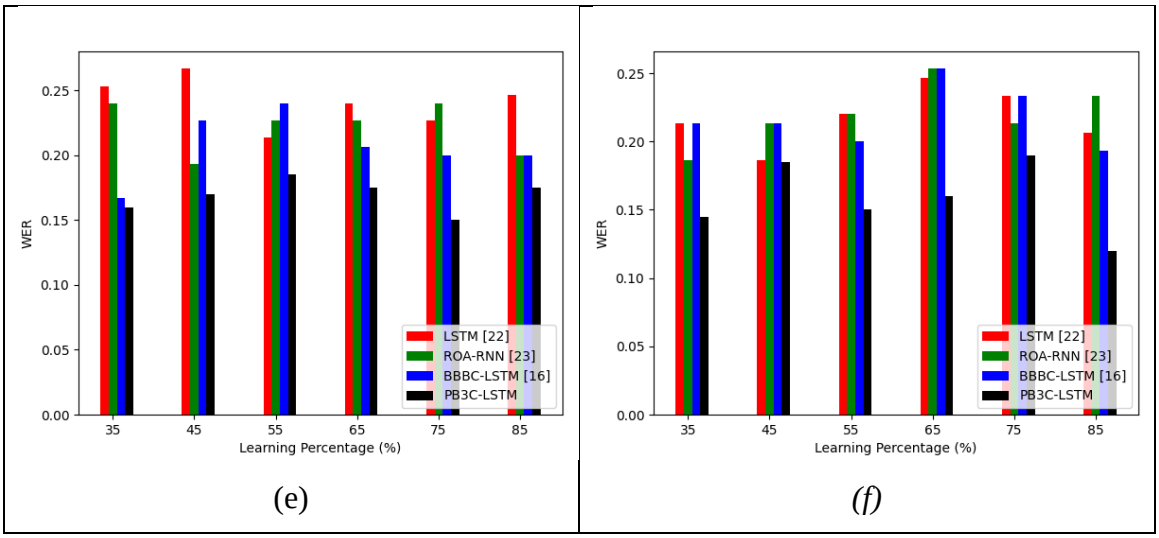


Fig. 6.4: Performance analysis of the proposed speech recognition model on Marathi language in terms of WER with different conventional approaches for "(a) Dataset 1, (b) Dataset 2, (c) Dataset 3, (d) Dataset 4, (e) Dataset 5 and (f) Dataset 6"

While examining dataset 4, the WER of the used PB3C-LSTM is 11 percent, 30.4 percent, and 36 percent higher than LSTM, ROA-RNN, and BB-BC-LSTM, respectively. While evaluating dataset 5, the WER of the used PB3C-LSTM is 37.5 percent, 40 percent, and 25 percent higher than LSTM, ROA-RNN, and BB-BC-LSTM, respectively. While examining dataset 6, the WER of the recommended PB3C-LSTM is 40.9 percent, 45.8 percent, and 31.5 percent higher than LSTM, ROA-RNN, and BB-BC-LSTM, respectively. As a result, when compared to other algorithms, the proposed voice recognition model using P3BC-LSTM outperforms them.

Overall performance analysis

Error actions such as WER, WAR, and SER, as shown in Table I, are used to evaluate the performance of the built speech recognition model. In terms of data set 1, the WER of the recommended PB 3C-LSTM is enhanced by 4.07%, 3.18%, and 3.57% as the LSTM, the ROA-RNN, and BB-BC-LSTM. WAR is of the proposed PB3C-LSTM, is improved by 1.7% , 5%, and 2.5% than LSTM, ROA-RNN, and BB-BC-LSTM, when the data set 2 is taken into account, respectively. The SER of the BB-LST MC is 2.29%, 4.05% and 2.79% higher than that of LSTM, BB-BC-LSTM, and ROA-RNN, respectively, when taking into account, the data set 5. The overall performance of the developed P3BC-LSTM speech recognition model is better than other algorithms used here.

6.4 Summary

This study used P3BC-LSTM to create a novel voice recognition model for the Marathi language. Four major procedures are included: "preprocessing, feature extraction, feature selection, and classification. Smoothing and median filtering procedures were used to prepare the voice samples for feature extraction. MFCC and spectral-based features were used to complete the task. Additionally, these features were narrowed down by PCA before moving on to classification with P3BC-LSTM. Using the P3BC algorithm, the P3BC-LSTM was proposed to optimise the number of hidden neurons and weights in order to provide speech signals that could be identified. Because of this, the suggested speech recognition model utilising PB3C-LSTM had a Word Accuracy Rate (WAR) that was 1.34 percent higher than LSTM and 3.34 percent higher than BB-BC-LSTM when using the Speech corpus. LSTM and BB-BC-LSTM have both achieved WER reductions of 4 and 6 percent, and SER reductions of 4.45 and 3.25 percent, respectively, with the suggested P3BC-LSTM model, and the findings are similar with the entire speech corpus.

CHAPTER 7: COMPARISON OF ALL PROPOSED APPROACHES

In this research work, three approaches are proposed for developing the speech recognition system for the Marathi language. The performances of these approaches are measured with the six measures – Accuracy, Sensitivity, Specificity, FPR, FNR, and MCC. All the experimental values are given in tables starting from Table 7.1 to Table 7.6.

Table 7.1 : Overall Performance Analysis for Speech Corpus 1

Measures	Accuracy	Sensitivity	Specificity	FPR	FNR	MCC
RCBO-DBN	'0.6713	'0.9862	'0.6649	'0.3350	0.0137	0.1913
GOA-RNN	0.9612	0.7946	0.9681	0.0318	0.2053	0.6158
PB3C-LSTM	0.972563	0.991977	0.503726	0.049627	0.008023	0.58987

Table 7.2 : Overall Performance Analysis for Speech Corpus 2

Measures	Accuracy	Sensitivity	Specificity	FPR	FNR	MCC
RCBO-DBN	0.7481	0.9935	0.7431	0.2568	0.0064	0.2318
GOA-RNN	0.9603	0.7823	0.9675	0.0324	0.2176	0.6048
PB3C- LSTM	0.972267	0.990881	0.504644	0.049536	0.009119	0.575444

Table 7.3 : Overall Performance Analysis for Speech Corpus 3

Measures	Accuracy	Sensitivity	Specificity	FPR	FNR	MCC
RCBO-DBN	0.6820	0.9930	0.6757	0.3242	0.0069	0.1979
GOA-RNN	0.9610	0.7861	0.0318	0.2138	0.2138	0.6083
PB3C- LSTM	0.973511	0.992843	0.503003	0.049697	0.007157	0.598597

Table 7.4 : Overall Performance Analysis for Speech Corpus 4

Measures	Accuracy	Sensitivity	Specificity	FPR	FNR	MCC
RCBO-DBN	0.5721	0.9795	0.5638	0.4361	0.0204	0.1530
GOA-RNN	0.9605	0.7877	0.9675	0.0324	0.2122	0.6073
PB3C- LSTM	0.975881	0.994637	0.509202	0.049079	0.005363	0.62404

Table 7.5 : Overall Performance Analysis for Speech Corpus 5

Measures	Accuracy	Sensitivity	Specificity	FPR	FNR	MCC
RCBO-DBN	0.7399	0.9895	0.7348	0.2651	0.0104	0.2259
GOA-RNN	0.9606	0.7771	0.9681	0.0318	0.2228	0.6053
PB3C- LSTM	0.977896	0.997286	0.502269	0.049773	0.002714	0.656609

Table 7.6 : Overall Performance Analysis for Speech Corpus 6

Measures	Accuracy	Sensitivity	Specificity	FPR	FNR	MCC
RCBO-DBN	0.6021	0.9595	0.5938	0.4151	0.0174	0.1930
GOA-RNN	0.9615	0.8110	0.9675	0.0324	0.1889	0.6180
PB3C- LSTM	0.968533	0.987609	0.495413	0.05045	0.012391	0.536977

The performance measure values are calculated and compared for all six corpora. After analyzing the parameter values given in the above tables, it is observed that PB3C-LSTM is the best suitable approach for the Marathi Speech Recognition development compared to RCBO-DBN and GOA-RNN proposed approaches.

CHAPTER 8: CONCLUSION

“Day-by-day communications are carried out in a multilingual nation such as India, which includes Hindi and English and in a regional language 17 such languages as Assamese, Tamil, Malayalam Gujarati, Telugu, Oriya, Urdu, Bengali, Sinkri, Kashmiri, Punjabi, Konkani, Marathi, Manipuri, Kannada, and Nepali”. A system isn't useful unless we develop SRS for regional languages considering the dialects of that language. Therefore, SRS development for regional languages is essential to convey technology to the masses. Much has been accomplished in Hindi, Tamil, and Bengali languages. It is noted that there is not so much more emphasis given to work in Marathi.

Chapter 1 discusses the fundamentals of the Natural Language Processing and Speech Recognition system. Chapter 2 describes the review of the literature with respect to various parameters such as technology and Indian languages research work. This chapter also includes studies on the Speech Recognition System Development Approaches, Speech Recognition Systems for Indian languages, Speech Recognition Challenges, Research Gap, Research Motivation, and Research Objectives.

The research tools and methods used for the study are explained in chapter 3. The research showed the instruments that were utilized for the study and the optimal strategy for the study. The researcher also gave extensive guidance on the selection of appropriate study equipment. The speech corpus is collected from the Center for Proliferation and Development of Indian Languages, Government of India.

This research proposed 4 new approaches for Marathi Speech Recognition System –

- Hybrid MFCC-SF feature extraction approach
- RCBO-DBN approach for Marathi Speech Recognition
- GOA-RNN approach for Marathi Speech Recognition
- PB3C-LSTM approach for Marathi Speech Recognition

The performance of these approaches was evaluated on the Speech Corpus collected from Indian Language Proliferation and Development Centre, Government of India. All these approaches are implemented using Python.

Chapter 4 analyzes various approaches performances for developing Marathi SRS and also proposes a new Duo feature-based RCBO-Deep Belief Network approach. The observations for RCBO-DBN are as follows -

- After performance analysis, it was observed that duo-feature performs better with the advantage of 12.7% than MFCC and 13% than Spectral features.
- DBN values almost all the factors used for quantifying the performance of the recognition compared to four other pattern recognition approaches.
- DBN gives about 81% accuracy which is higher than the approaches used for the comparison such as SVM, KNN, ANN, DNN.
- The DBN algorithm is too costly in terms of computation because of its architecture.

Chapter 5 proposes another approach GOA-Recurrent Neural Network Marathi SRS. The observations found after implementing this approach are as follows –

- The accuracy of the proposed GOA-RNN model was 5.2%, 1.16%, and 0.86% progressed than RNN, LSTM, and Res-CNN, respectively for speech corpus 1.
- The WER of the proposed GOA-RNN model was 3.84%, 1.06%, and 0.79% improved than RNN, LSTM, and CNN, respectively, for speech corpus one, and it has similar results with the remaining speech corpus.
- The RNN is good for Natural Language Processing but due to its short-term memory, it is not suitable for long statements and predicting the next word. It can be used for recognizing small sentences or words.

The six chapter proposes PB3C-LSTM Marathi SRS Approach. The observations for PB3C-LSTM are as follows –

- Therefore, from the experimental results, the Word Accuracy Rate (WAR) of the proposed speech recognition model using PB3C-LSTM was 1.34% and 3.34%

increased than LSTM and BB-BC-LSTM, respectively, while considering the Speech corpus one.

- The WER of the proposed P3BC-LSTM model has attained 4% and 6% less WER, and 4.45% and 3.25% less SER than LSTM and BB-BC-LSTM, respectively, for speech corpus one, and it has similar results with remaining speech corpus.
- The LSTM approach is an extension to the RNN with the addition of Long- term memory. This algorithm is best suitable for Speech Recognition System, and we can use this for deploying the Marathi Speech Recognition System model.

The seventh chapter compares and analyzes the proposed approach performances with the given set of parameters. Finally, the research suggests that using PB3C-LSTM for developing the Marathi SRS is a suitable approach.

The common observations are given below –

- Very rare research has been done on Indian Language Speech Recognition Systems. Especially in the Marathi language, there was significantly less research done with limited data, conventional techniques, and not considering the various characteristics of Marathi Speech.
- If one system architecture is working well for the Hindi language, we cannot guarantee the same framework will perform the same due to various reasons.
- The heuristic algorithms are used with the neural network to avoid the local minima problem caused by standard optimizers such as Gradient Descent, AdaGrad, AdaDelta, Stochastic Gradient, etc.

8.1 Future Scope

This work can be extended to design the Marathi Speech Recognition system for children. It will be an interesting task to create a child speech corpus with different characteristics to make that system with minimum WER and SER. This will be helpful in designing smart home appliances and education products having a voice as an interface.

PUBLICATIONS

1. “Automatic Speech Recognition Systems for Regional Languages in India” published in "International Journal of Recent Technology and Engineering (TM)", ISSN: 2277-3878, Volume-8, Issue-2S3, July 2019 [Scopus Indexed Journal]

   https://www.ijrte.org/wpcontent/uploads/papers/v8i2S3/B11080782S319.pdf

2. “Acquaintance with Natural Language Processing for Smart Society” Submitted and accepted in 6th International Conference Energy & City of the future, EVF2019 [Scopus Indexed Journal - Web of Conferences] [Presented].

   https://www.e3sconferences.org/articles/e3sconf/abs/2020/30/e3sconf_evf2020_2006/e3sconf_evf2020_02006.html

3. “Comparing different pattern recognition approaches for Building Marathi ASR System” published in International Journal of Advanced Science and Technology, Vol. 29, No. 5, (2020), pp. 4615 – 4623. [Scopus Indexed Journal]

   https://www.researchgate.net/publication/341592095_Comparing_Different_Pattern_Recognition_Approaches_of_Building_Marathi_ASR_System

4. “Duo Features with Hybrid-Meta-Heuristic-Deep Belief Network based Pattern Recognition for Marathi Speech Recognition” (Springer Conference)
   https://link.springer.com/chapter/10.1007/978-981-16-3346-1_53

5. “Enhanced Marathi Speech Recognition enabled by Grasshopper Optimisation-based Recurrent Neural Network” accepted in Computer Systems Science and Engineering (SCI and Scopus Indexed Journal)

6. “Parallel Big Bang-Big Crunch-LSTM Approach for developing a Marathi Speech Recognition System” accepted in Intelligent Automation & Soft Computing (SCIE and Scopus Indexed)

Copyrights/ Patent

1. Copyright filled on “Duo Features With Hybrid Metaheuristics Deep Belief Network-Based Marathi Speech Recognition System”, diary No: 224/2021-CO/SW

2. Copyright filled on Enhanced Marathi Speech Recognition enabled by Grasshopper optimization-based Recurrent Neural Network Dairy No: 16443/2021-CO/SW

3. Patent filed on topic – “Educational Toy for Marathi Speaking Children below 5 years age”, Application Number 202111046120

REFERENCES

[1] R. P. Bachate and A. Sharma, “Acquaintance with Natural Language Processing for Building Smart Society,” *E3S Web Conf.*, vol. 170, p. 02006, 2020.

[2] X. Huang and L. Deng, “An overview of modern speech recognition,” *Handb. Nat. Lang. Process. Second Ed.*, pp. 339–366, 2010.

[3] and H. W. H. X. Huang, A. Acero, *Spoken Language Processing*, vol. 148. 2001.

[4] R. P. Bachate and A. Sharma, “Comparing different pattern recognition approaches of building marathi asr system,” *Int. J. Adv. Sci. Technol.*, vol. 29, no. 5, pp. 4615–4623, 2020.

[5] D. Jurafsky and D. Martin, *Speech and Language Processing: An introduction to natural language processing*. 2020.

[6] J. Li, L. Deng, Y. Gong, S. Member, R. Haeb-umbach, and S. Member, “An Overview of Noise-Robust Automatic Speech Recognition,” vol. 22, no. 4, pp. 745–777, 2014.

[7] S. B. DAVIS and P. MERMELSTEIN, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences,” *Readings Speech Recognit.*, no. 4, pp. 65–74, 1990.

[8] Lawrence Rabiner and Biing Hwang Juang, *Fundamentals of speech recognition*. 2014.

[9] M. A. Anusuya and S. K. Katti, “Speech Recognition by Machine: A Review,” *IJCSIS) Int. J. Comput. Sci. Inf. Secur.*, vol. 6, no. 3, pp. 181–205, 2009.

[10] HIROAKI SAKOE and SEIBI CHIBA, “Dynamic Programming Algorithm Optimization for Spoken Word Recognition,” *IEEE Trans. Acoust.*, vol. VOL. ASSP-, no. 1, pp. 43–49, 1978.

[11] J. A. Bilmes, “What HMMs can do,” *IEICE Trans. Inf. Syst.*, vol. E89-D, no. 3, pp. 869–891, 2006.

[12] N. Morgan *et al.*, “Pushing the envelope - Aside,” *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 81–88, 2005.

[13] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition*, no. June. 1994.

[14] N. D. Smith and M. J. F. Gales, “Using SVMS and discriminative models for speech recognition,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 1, pp. 77–80, 2002.

[15] R. P. Bachate and A. Sharma, “Automatic speech recognition systems for regional languages in India,” *Int. J. Recent Technol. Eng.*, vol. 8, no. 2 Special Issue 3, pp. 585–592, Jul. 2019.

[16] A. Becerra, J. I. de la Rosa, and E. González, “Speech recognition in a dialog system: from conventional to deep processing: A case study applied to Spanish,” *Multimed. Tools Appl.*, vol. 77, no. 12, pp. 15875–15911, 2018.

[17] S. Gaikwad, B. Gawali, and S. Mehrotra, “Creation of Marathi speech corpus for automatic speech recognition,” *2013 Int. Conf. Orient. COCOSDA Held Jointly with 2013 Conf. Asian Spok. Lang. Res. Eval. O-COCOSDA/CASLRE 2013*, no. November, 2013.

[18] V. Bhardwaj and V. Kukreja, “Effect of pitch enhancement in Punjabi children’s speech recognition system under disparate acoustic conditions,” *Appl. Acoust.*, vol. 177, p. 107918, 2021.

[19] P. Upadhyaya, O. Farooq, M. R. Abidi, and Y. V. Varshney, “Continuous Hindi speech recognition model based on Kaldi ASR toolkit,” in *Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2017*, 2018, vol. 2018-Janua, no. 0, pp. 786–789.

[20] S. Toshniwal *et al.*, “MULTILINGUAL SPEECH RECOGNITION WITH A SINGLE END-TO-END MODEL Toyota Technological Institute at Chicago,” pp. 4904–4908, 2018.

[21] A. Narkhede, “Efficient Method for Isolated Marathi Digits Recognition using DWT and Soft Computing Techniques,” *2018 3rd Int. Conf. Internet Things Smart Innov. Usages*, pp. 1–5, 2018.

[22] N. D. Londhe, “Recognition for Chhattisgarhi,” *2018 5th Int. Conf. Signal Process. Integr. Networks*, pp. 667–671, 2018.

[23] N. D. Londhe and G. B. Kshirsagar, “Chhattisgarhi speech corpus for research and development in automatic speech recognition,” *Int. J. Speech Technol.*, vol. 21, no. 2, pp. 193–210, 2018.

[24] P. Sahu, M. Dua, and A. Kumar, “Challenges and Issues in Adopting Speech Recognition,” *Springer*, vol. 664, pp. 209–215, 2017.

[25] S. J. Arora, “Automatic Speech Recognition : A Review Automatic Speech Recognition : A Review,” no. September, pp. 33–44, 2017.

[26] S. K. Saksamudre and R. R. Deshmukh, “A Review on Different Approaches for Speech Recognition System,” no. September, 2015.

[27] L. Deng, G. Hinton, and B. Kingsbury, “New types of deep neural network learning for speech recognition and related applications: An overview,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 8599–8603, 2013.

[28] N. Chadha, R. C. Gangwar, and R. Bedi, “Current Challenges and Application of Speech Recognition Process using Natural Language Processing: A Survey,” *Int. J. Comput. Appl.*, vol. 131, no. 11, pp. 28–31, 2015.

[29] H. Petkar, “A Review of Challenges in Automatic Speech Recognition,” *Int. J. Comput. Appl.*, vol. 151, no. 3, pp. 23–26, 2016.

[30] J. Ramirez, J. M., and J. C., “Voice Activity Detection. Fundamentals and Speech Recognition System Robustness,” *Robust Speech Recognit. Underst.*, no. June, 2007.

[31] J. Guo *et al.*, “Deep neural network based i-vector mapping for speaker verification using short utterances,” *Speech Commun.*, 2018.

[32] M. Li, D. Tang, J. Zeng, T. Zhou, and H. Zhu, “An Automated Assessment Framework for A typical Prosody and Stereotyped Idiosyncratic Phrases related to Autism Spectrum Disorder,” *Comput. Speech Lang.*, 2018.

[33] H. O. ; H Belar, “Phonetic Typewriter,” *IRE Trans. Audio*, vol. 5, no. 4, pp. 90–95, 1957.

[34] S. Huang and S. Renals, “Hierarchical Bayesian language models for conversational speech recognition,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 8, pp. 1941–1954, 2010.

[35] S. Supriya and S. M. Handore, “Speech recognition using HTK toolkit for Marathi language,” in *IEEE International Conference on Power, Control, Signals and Instrumentation Engineering, ICPCSI 2017*, 2018, pp. 1591–1597.

[36] S. Sun, B. Zhang, L. Xie, and Y. Zhang, “An unsupervised deep domain adaptation

approach for robust speech recognition," *Neurocomputing*, vol. 257, pp. 79–87, 2017.

[37] S. Hazmoune, F. Bougamouza, S. Mazouzi, and M. Benmohammed, "A new hybrid framework based on Hidden Markov models and K-nearest neighbors for speech recognition," *Int. J. Speech Technol.*, vol. 21, no. 3, pp. 689–704, 2018.

[38] J. Du, Q. Wang, T. Gao, Y. Xu, L. Dai, and C. H. Lee, "Robust speech recognition with speech enhanced deep neural networks," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, no. September, pp. 616–620, 2014.

[39] R. Subhashini and V. J. Senthil Kumar, "A framework for efficient information retrieval using NLP techniques," *Commun. Comput. Inf. Sci.*, vol. 142 CCIS, pp. 391–393, 2011.

[40] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson, E. Saltzman, and M. Tiede, "Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition," *Speech Commun.*, vol. 89, pp. 103–112, 2017.

[41] M. Lee, J. Lee, and J. Chang, "Ensemble of jointly trained deep neural network-based acoustic models for reverberant speech recognition," *Digit. Signal Process.*, vol. 85, pp. 1–9, 2019.

[42] T. Bhowmik, S. Kumar, and D. Mandal, "Manner of articulation based Bengali phoneme classification," *Int. J. Speech Technol.*, vol. 21, no. 2, pp. 233–250, 2018.

[43] T. Zia and U. Zahid, "Long short-term memory recurrent neural network architectures for Urdu acoustic modeling," *Int. J. Speech Technol.*, vol. 22, no. 1, pp. 21–30, 2019.

[44] T. G. Y. H. S. Jayanna, "A spoken query system for the agricultural commodity prices and weather information access in Kannada language," *Int. J. Speech Technol.*, vol. 20, no. 3, pp. 635–644, 2017.

[45] P. Das, K. Acharjee, P. Das, and V. Prasad, "VOICE RECOGNITION SYSTEM : SPEECH-TO-TEXT," no. July, 2016.

[46] S. F. Worgan and R. K. Moore, "Towards the detection of social dominance in dialogue," *Speech Commun.*, vol. 53, no. 9–10, pp. 1104–1114, 2011.

[47] J. H. and D. B., "Speech Recognition System Architecture for Gujarati Language," *Int. J. Comput. Appl.*, vol. 138, no. 12, pp. 28–31, 2016.

[48] U. G. Patil, S. D. Shirbahadurkar, and A. N. Paithane, “Automatic Speech Recognition of isolated words in Hindi language using MFCC,” *Int. Conf. Comput. Anal. Secur. Trends, CAST 2016*, pp. 433–438, 2017.

[49] K. Mannepalli, P. N. Sastry, and M. Suman, “MFCC-GMM based accent recognition system for Telugu speech signals,” *Int. J. Speech Technol.*, vol. 19, no. 1, pp. 87–93, 2016.

[50] R. K. Aggarwal and M. Dave, “Integration of multiple acoustic and language models for improved Hindi speech recognition system,” *Int. J. Speech Technol.*, vol. 15, no. 2, pp. 165–180, 2012.

[51] M. Kalamani, M. Krishnamoorthi, and R. S. Valarmathi, “Continuous Tamil Speech Recognition technique under non stationary noisy environments,” *Int. J. Speech Technol.*, vol. 22, no. 1, pp. 47–58, 2019.

[52] S. Sruba, B. Sanjib, and K. Kalita, “Speech recognition with reference to Assamese language using novel fusion technique,” *Int. J. Speech Technol.*, vol. 0, no. 0, p. 0, 2018.

[53] J. Guglani and A. N. Mishra, “Continuous Punjabi speech recognition model based on Kaldi ASR toolkit,” *Int. J. Speech Technol.*, vol. 21, no. 2, pp. 211–216, 2018.

[54] V. Kadyan, A. Mantri, R. K. Aggarwal, and A. Singh, “A comparative study of deep neural network based Punjabi-ASR system,” *Int. J. Speech Technol.*, vol. 22, no. 1, pp. 111–119, 2019.

[55] E. Yılmaz, M. Mclaren, H. Van Den Heuvel, D. A. Van Leeuwen, E. Yılmaz, and M. Mclaren, “Semi-supervised acoustic model training for speech,” *Speech Commun.*, 2018.

[56] V. D. W. Dx and R. N. Dx, “Synthesised bigrams using word embeddings for code-switched ASR of four South African language pairs,” *Comput. Speech Lang.*, 2018.

[57] M. El, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition : Features , classification schemes , and databases,” *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

[58] G. Hinton *et al.*, “Deep Neural Networks for Acoustic Modeling in Speech Recognition,” pp. 1–27.

[59] I. Kauppinen and K. Roth, “Improved Noise Reduction in Audio Signals Using

Spectral Resolution Enhancement With Time-Domain Signal Extrapolation," vol. 13, no. 6, pp. 1210–1216, 2005.

[60] T. Y. G, "Development and Comparison of ASR Models using Kaldi for Noisy and Enhanced Kannada Speech Data," pp. 1832–1838, 2017.

[61] D. Malewadi, "Development of Speech recognition technique for Marathi numerals using MFCC & LFZI algorithm."

[62] R. V. Darekar and A. P. Dhande, "Emotion recognition from Marathi speech database using adaptive artificial neural network," *Biol. Inspired Cogn. Archit.*, vol. 23, no. November 2017, pp. 35–42, 2018.

[63] S. P. Patil and S. L. Lahudkar, "Hidden-Markov-model based statistical parametric speech synthesis for Marathi with optimal number of hidden states," *Int. J. Speech Technol.*, vol. 22, no. 1, pp. 93–98, 2019.

[64] S. Najnin and B. Banerjee, "Speech recognition using cepstral articulatory features," *Speech Commun.*, vol. 107, no. December 2018, pp. 26–37, 2019.

[65] A. H. Abdelaziz, "Comparing Fusion Models for DNN-Based Audiovisual Continuous Speech Recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 3, pp. 475–484, 2018.

[66] F. Tao and C. Busso, "Gating Neural Network for Large Vocabulary Audiovisual Speech Recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 7, pp. 1286–1298, 2018.

[67] J. Deng, X. Xu, Z. Zhang, S. Fruhholz, and B. Schuller, "Semisupervised Autoencoders for Speech Emotion Recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 1, pp. 31–43, 2018.

[68] Y. Shi, J. Bai, P. Xue, and D. Shi, "Fusion Feature Extraction Based on Auditory and Energy for Noise-Robust Speech Recognition," *IEEE Access*, vol. 7, pp. 81911–81922, 2019.

[69] P. Sharma, V. Abrol, and A. K. Sao, "Deep-Sparse-Representation-Based Features for Speech Recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 11, pp. 2162–2175, 2017.

[70] A. Choudhary, R. Chauhan, and G. Gupta, "Automatic Speech Recognition System for Isolated & Connected Words of Hindi Language By Using Hidden Markov

Model Toolkit (HTK)," *Proc. Int. Conf. Emerg. Trends Eng. Technol.*, pp. 847–853, 2013.

[71] Preeti Sain and Parneet Kau, "Automatic speech recognition: A Review," *ICEIS 2003 - Proc. 5th Int. Conf. Enterp. Inf. Syst.*, vol. 1, no. iii, pp. IS5–IS10, 2003.

[72] M. Moneykumar *et al.*, "Malayalam Word Identification for Speech," *Int. J. Eng. Sci.*, vol. 15, no. December, pp. 22–26, 2014.

[73] A. L. Maas *et al.*, "Building DNN acoustic models for large vocabulary speech recognition," *Comput. Speech Lang.*, vol. 41, pp. 195–213, 2017.

[74] T. Bhardwaj and P. Somvanshi, *Continuous Hindi Speech Recognition usin Kaldi ASR Based on Deep neural network*, vol. 748. Springer Singapore, 2019.

[75] R. Li, C. Yin, X. Zhang, and B. David, *Recent Developments in Intelligent Computing, Communication and Devices*, vol. 752. Springer Singapore, 2019.

[76] H. Kamper, A. Jansen, and S. Goldwater, "A segmental framework for fully-unsupervised large-vocabulary speech recognition," *Comput. Speech Lang.*, vol. 46, pp. 154–174, 2017.

[77] S. Xue, H. Jiang, L. Dai, and Q. Liu, "Speaker Adaptation of Hybrid NN/HMM Model for Speech Recognition Based on Singular Value Decomposition," *J. Signal Process. Syst.*, vol. 82, no. 2, pp. 175–185, 2016.

[78] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018.

[79] M. Peters *et al.*, "Deep Contextualized Word Representations," 2018, pp. 2227–2237.

[80] J. Gu, Y. Wang, Y. Chen, K. Cho, and V. O. K. Li, "Meta-Learning for Low-Resource Neural Machine Translation," *Comput. Lang.*, 2018.

[81] G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato, "Phrase-Based & Neural Unsupervised Machine Translation," 2018.

[82] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni, "What you can cram into a single vector: Probing sentence embeddings for linguistic properties," 2018.

[83] S. R. Jeremy Howard, "Universal Language Model Fine-tuning for Text Classification," *Proc. 56th Annu. Meet. Assoc. Comput. Linguist. (Long Pap.*, pp.

328–339, 2018.

[84] M. Barrett, J. Bingel, N. Hollenstein, M. Rei, and A. Søgaard, “Sequence Classification with Human Attention,” in *CoNLL*, 2018, no. CoNLL, pp. 302–312.

[85] P. Shubham, K. Pratik, P. Rahul, and A. K. Gupta, “Design and Development of Word Recognition for Marathi Language,” vol. 7, no. 5, pp. 2337–2340, 2016.

[86] P. Mittal and N. Singh, “Development and analysis of Punjabi ASR system for mobile phones under different acoustic models,” *Int. J. Speech Technol.*, vol. 0, no. 0, p. 0, 2019.

[87] S. G. Koolagudi, A. Bharadwaj, and Y. V. S. Murthy, “Dravidian language classification from speech signal using spectral and prosodic features,” *Int. J. Speech Technol.*, vol. 20, no. 4, pp. 1005–1016, 2017.

[88] P. Pal Singh, “An Approach to Extract Feature using MFCC,” *IOSR J. Eng.*, vol. 4, no. 8, pp. 21–25, 2014.

[89] and G. H. Abdel-rahman Mohamed, George E. Dahl, “Acoustic Modeling Using Deep Belief Networks,” *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 20, no. 1, 2012.

[90] P. Gurunath Shivakumar and P. Georgiou, “Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations,” *Comput. Speech Lang.*, vol. 63, 2020.

[91] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[92] A. M. and G. H. Alex Graves, “Speech Recognition with Deep Recurrent Neural Networks , Department of Computer Science, University of Toronto,” *Dep. Comput. Sci. Univ. Toronto*, vol. 3, no. 3, pp. 45–49, 2013.

[93] D. Dash, M. Kim, K. Teplansky, and J. Wang, “Automatic speech recognition with articulatory information and a unified dictionary for Hindi, Marathi, Bengali, and oriya,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2018-Septe, no. August, pp. 1046–1050, 2018.

[94] M. D. Prasetio, T. Hayashida, I. Nishizaki, and S. Sekizaki, “Deep belief network optimization in speech recognition,” *Proc. - 2017 Int. Conf. Sustain. Inf. Eng. Technol. SIET 2017*, vol. 2018-Janua, pp. 138–143, 2018.

[95] S. Paulose, S. Nath, and K. Samudravijaya, “Marathi Speech Recognition,” in *The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages 29-31 August 2018, Gurugram, India*, 2018, no. August, pp. 235–238.

[96] K. R. Ghule and R. R. Deshmukh, “Automatic Speech Recognition of Marathi isolated words using Neural Network,” vol. 6, no. 5, pp. 4296–4298, 2015.

[97] S. Lokesh, P. Malarvizhi Kumar, M. Ramya Devi, P. Parthasarathy, and C. Gokulnath, “An Automatic Tamil Speech Recognition system by using Bidirectional Recurrent Neural Network with Self-Organizing Map,” *Neural Comput. Appl.*, vol. 31, no. 5, pp. 1521–1531, 2019.

[98] K. Sangramsing, “Marathi Speech Recognition System Using Hidden Markov Model Toolkit,” *Open Acces*, no. June, 2015.

[99] R. P. Bachate, A. Sharma, and A. Singh, “Duo Features with Hybrid-Meta-Heuristic-Deep Belief Network Based Pattern Recognition for Marathi Speech Recognition,” in *Advances in Intelligent Systems and Computing, Springer, Singapore*, vol. vol 1374, 2022, pp. 665–673.

[100] S. Mirjalili, S. M. Mirjalili, and A. Lewis, “Grey Wolf Optimizer,” *Adv. Eng. Softw.*, vol. 69, pp. 46–61, 2014.

[101] S. Mirjalili and A. Lewis, “The Whale Optimization Algorithm,” *Adv. Eng. Softw.*, vol. 95, pp. 51–67, 2016.

[102] S. Saremi, S. Mirjalili, and A. Lewis, “Biogeography-based optimisation with chaos,” *Neural Comput. Appl.*, vol. 25, no. 5, pp. 1077–1097, 2014.

[103] D. Binu and B. S. Kariyappa, “RideNN: A New Rider Optimization Algorithm-Based Neural Network for Fault Diagnosis in Analog Circuits,” *IEEE Trans. Instrum. Meas.*, vol. 68, no. 1, pp. 2–26, 2019.

[104] K. Vaitheki, N. B. Arune Kumar, and K. Suresh Joseph, “A QoS-oriented novel optimization schemes for web service composition for improved healthcare,” *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 12, pp. 3442–3447, 2019.

[105] M. Yarlagadda, K. Gangadhara Rao, and A. Srikrishna, “Frequent itemset-based feature selection and Rider Moth Search Algorithm for document clustering,” *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, 2019.

[106] M. Swain, S. Sahoo, A. Routray, P. Kabisatpathy, and J. N. Kundu, “Study of feature

combination using HMM and SVM for multilingual Odiya speech emotion recognition," *Int. J. Speech Technol.*, vol. 18, no. 3, 2015.

[107] S. Huang and S. Renals, "Hierarchical Bayesian language models for conversational speech recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 8, pp. 1941–1954, 2010.

[108] Y. He and X. Dong, "Real time speech recognition algorithm on embedded system based on continuous Markov model," *Microprocess. Microsyst.*, vol. 75, p. 103058, 2020.

[109] A. Mohan, R. Rose, S. H. Ghalehjegh, and S. Umesh, "Acoustic modelling for speech recognition in Indian languages in an agricultural commodities task domain," *Speech Commun.*, vol. 56, no. 1, pp. 167–180, 2014.

[110] G. Pironkov, S. U. Wood, and S. Dupont, "Hybrid-task learning for robust automatic speech recognition," *Comput. Speech Lang.*, vol. 64, 2020.

[111] G. S. V. S. Sivaram, S. K. Nemala, N. Mesgarani, and H. Hermansky, "Data-driven and feedback based spectro-temporal features for speech recognition," *IEEE Signal Process. Lett.*, vol. 17, no. 11, pp. 957–960, 2010.

[112] J. Guglani and A. N. Mishra, "Automatic speech recognition system with pitch dependent features for Punjabi language on KALDI toolkit," *Appl. Acoust.*, vol. 167, p. 107386, 2020.

[113] P. Smit, S. Virpioja, and M. Kurimo, "Advances in subword-based HMM-DNN speech recognition across languages," *Comput. Speech Lang.*, vol. 66, p. 101158, 2021.

[114] M. A. J. Sathya and S. P. Victor, "Noise Reduction Techniques and Algorithms For Speech Signal Processing," *Int. J. Sci. Eng. Res.*, vol. 6, no. 1, p. 2015, 2015.

[115] S. Saremi, S. Mirjalili, and A. Lewis, "Grasshopper Optimisation Algorithm: Theory and application," *Adv. Eng. Softw.*, vol. 105, pp. 30–47, 2017.

[116] S. Deena, M. Hasan, M. Doulaty, O. Saz, and T. Hain, "Recurrent neural network language model adaptation for multi-genre broadcast speech recognition and alignment," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 3, pp. 572–582, 2019.

[117] V. Passricha and R. Kumar, "Convolutional support vector machines for speech

recognition," *Int. J. Speech Technol.*, vol. 0, no. 0, p. 0, 2018.

[118] Y. H. Tu *et al.*, "An iterative mask estimation approach to deep learning based multi-channel speech recognition," *Speech Commun.*, vol. 106, no. November 2018, pp. 31–43, 2019.

[119] I. S. Kipyatkova and A. A. Karpov, "A study of neural network Russian language models for automatic continuous speech recognition systems," *Autom. Remote Control*, vol. 78, no. 5, pp. 858–867, 2017.

[120] P. Zhou, H. Jiang, L. R. Dai, Y. Hu, and Q. F. Liu, "State-Clustering Based Multiple Deep Neural Networks Modeling Approach for Speech Recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 23, no. 4, pp. 631–642, 2015.

[121] S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and Q. Liu, "Fast adaptation of deep neural network based on discriminant codes for speech recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 1713–1725, 2014.

[122] S. Kumar, A. Singh, and S. Walia, "Parallel Big Bang–Big Crunch Global Optimization Algorithm: Performance and its Applications to routing in WMNs," *Wirel. Pers. Commun.*, vol. 100, no. 4, pp. 1601–1618, 2018.

[123] O. K. Erol and I. Eksin, "A new optimization method: Big Bang-Big Crunch," *Adv. Eng. Softw.*, vol. 37, no. 2, pp. 106–111, 2006.

[124] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech Emotion Classification Using Attention-Based LSTM," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 11, pp. 1675–1685, 2019.

[125] R. Venkateswarlu, R. Vasantha Kumari, and G. V. JayaSri, "Speech Recognition By Using Recurrent Neural Networks," *Int. J. Sci. Eng. Res.*, vol. 2, no. 6, pp. 1–7, 2011.