Chapter 1

Introduction

1.1 Speech Recognition

Auditory perception is one of the crucial abilities that a human possesses as a part of human intelligence, along with vision, speech, critical thinking, and taking decisions. As humans, we can hear the speech of other humans, the sounds of nature, the noise of machines, the melody of music, and songs. In the context of human intelligence, speech recognition is the ability of a human to hear and understand who is speaking, which language is being spoken, what is being said, and the meaning of it. In contrast to it, artificial intelligence, or machine intelligence, is a process of mimicking human intelligence of auditory perception by a machine. Hence, speech recognition by a machine mimics the auditory perception of humans.

The main areas of artificial intelligence are computer vision, speech synthesis, speech recognition, natural language processing, planning, and decision-making. Automatic speech recognition is a part of artificial intelligence in which a machine is trained to identify who is speaking, what is being said, and comprehend its meaning. It is a procedure for designing an intelligent machine that can automatically recognise natural human speech using the information contained in the digital speech signal. According to Rabiner [1], researchers and scientists have been trying to design an intelligent machine that can recognise the spoken word and understand its meaning for many decades, but we are still far from developing a machine that can comprehend spoken discourse on any subject from all speakers in all settings and languages.

The main reason for this is that the process is challenging because of its interdisciplinary nature [1]. Research in this area requires knowledge of signal processing, physics (acoustics), pattern recognition, communication and information theory, linguistics, physiology, computer science, psychology, and applied mathematics [1]. One more challenge is that

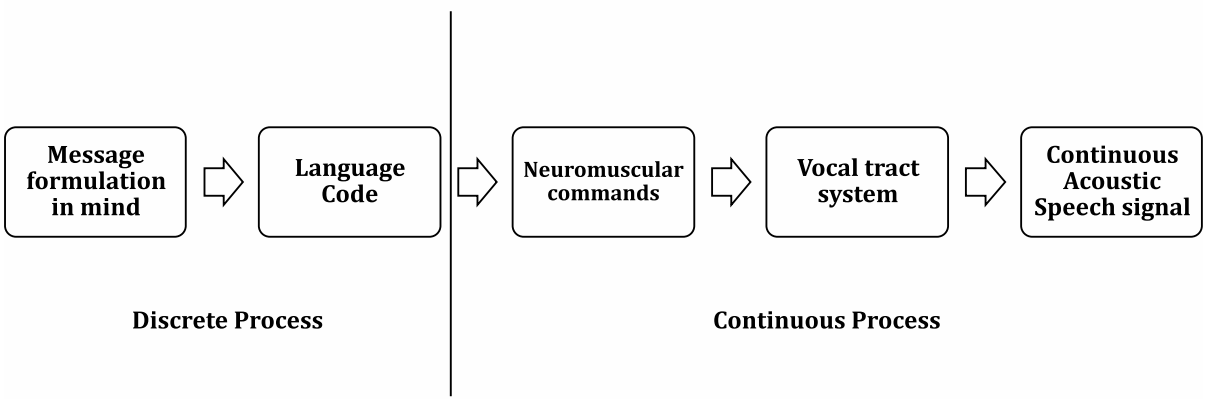

Figure 1.1: Speech production process [1]

the recorded or otherwise real-time speech signals include the surrounding noise. Moreover, there may be variations in pronunciations, and accents also differ over geography. Hence, more research is required in this field to make an accurate automatic speech recognition system that can recognise natural human speech without any error [5]. To understand the process of speech recognition, let us first understand speech production and the speech perception process.

1.2 Speech Production Process

Human speech production is a process of speaking by mouth. But it all starts with the formulation of a message in the human mind. Here, language is the first thing selected by the mind. The message is converted into language code in terms of a set of phonemes, which are the smallest unit of speech sound. The process was discrete up to this point, but now it is continuous. Then, neuromuscular commands are executed from the human brain, which determines the shape of the vocal tract and the vocal chords that vibrate to produce sound. The neuromuscular commands also determine the duration, loudness, and pitch of speech. Finally, a speech signal with different phoneme sounds is produced, depending on the shape of the vocal tract [1]. Diagrammatically, this process is summarised in the Figure 1.1.

The most important organ for speech production is the vocal tract. The vocal tract consists of various parts, starting from the vocal cord and up to the lips. The average length of the vocal tract is about 17 cm. It has a varying cross-sectional area from 0 to 20 cm^2. It takes different shapes based on the position of the tongue, lips, jaws, and velum. The combination of all this gives different shapes to the vocal tract, which produces different sounds. For the production of some phonemes, nasal activity also takes place. This involves the nasal tract. It starts from the velum and ends at the nose. The nasal tract

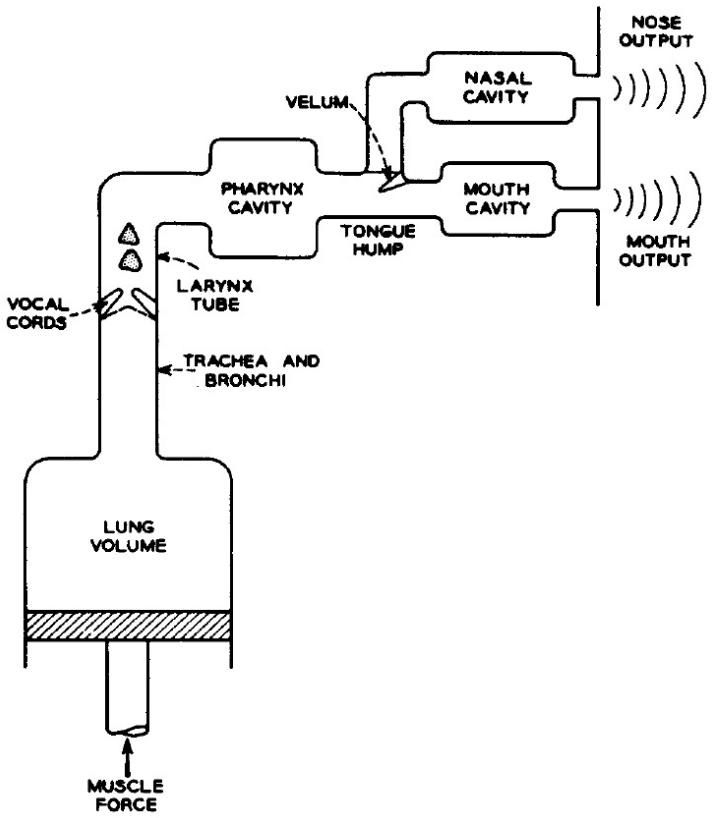

Figure 1.2: Schematic diagram of nasal tract and vocal tract [1]

and vocal tract can work together to produce nasal sound. This happens when the velum is lowered [1]. The vocal tract and nasal tract are summarised by a schematic diagram shown in the Figure 1.2.

1.3 Human and Machine Speech Perception

To understand automatic speech recognition technology in terms of artificial intelligence, first we need to understand how natural human intelligence plays a role in natural speech perception, done by the combination of acts of the human ear and brain, as explained in the following section.

1.3.1 Human Speech Perception

Consider that person A is speaking and person B is listening. In this example, when person A is speaking, the speech signal is generated and propagates towards person B. First, the acoustic signal is processed along the basilar membrane in the inner ear of person B, as shown in the Figure 1.3. The basilar membrane is a stiff structural element within the cochlea of the inner ear, as shown in the Figure 1.4. This acts like a dynamic

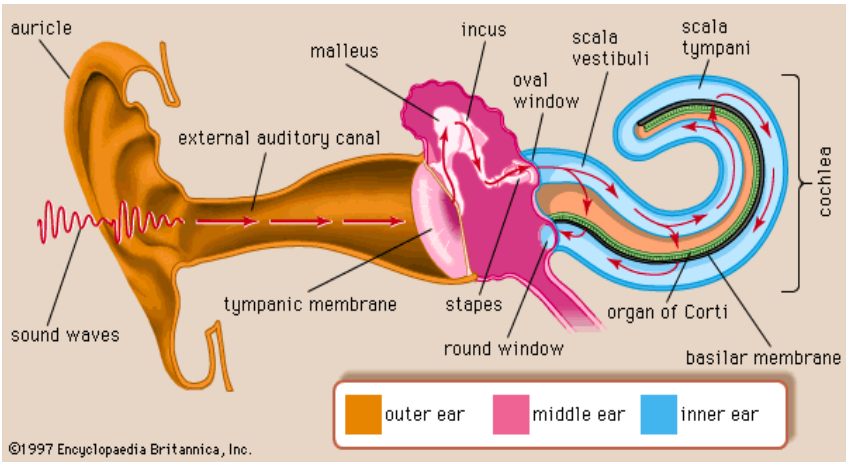

Figure 1.3: Expanded view of middle and inner ear [2]

spectral analyser of the incoming speech signal [1]. Further, the cochlea is connected with the auditory nerve. At the output of the basilar membrane, a neural transduction process converts the spectral signal into activity signals in the auditory nerve as shown in the Figure 1.4. This process corresponds to the feature extraction of speech signals [1]. The auditory nerve is further connected with neurons in the brain. Neural activity along the auditory nerve is converted into a language code in the neurons of the brain [1]. This corresponds to the classification of the extracted features. This is how a speech by person A is comprehended by person B. The goal of this research is to train a machine that imitates this speech perception process.

1.3.2 Machine Speech Perception

From the above example, we can understand that there are two main steps involved in speech recognition by machine:

1. Feature extraction from the digital speech signal.
2. Training a model with labelled features that can classify the spoken content.

Then this model can be useful to identify features in an unknown speech, and this is how a speech is recognised by a machine. Various methods of feature extraction and model training, along with their comparison, are explored in this thesis. The input to the machine speech recognition process is a recorded speech signal.

1.4 States of Speech Signal

A speech signal is a signal consisting of various types of states. Mainly, it involves the following three states [1]:

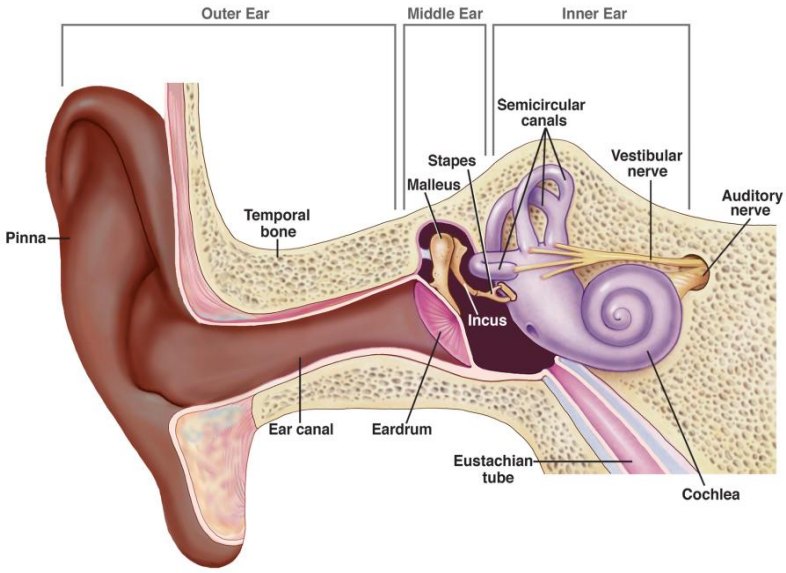

Figure 1.4: Physiological model of the human ear [3]

1. **Silences state**: The silence state represents a part of the speech signal in which there is no sound.

2. **Voiced state**: The voiced state represents a part of the speech signal in which the speech sounds can be heard. This happens when air flowing from the lungs leads to vibrations in the vocal chords. This generates a fundamental frequency, which is nothing but a periodic triangular-shaped pulse trail. The voiced state has a frequency of 80-350 Hz.

3. **Unvoiced state**: The unvoiced state represents a part of the speech signal during which the vocal chords are not vibrated. The air passes through the narrow passage, resulting in turbulence. This is defined as noise in the speech signal.

The shape of the speech signal depends on the shape and narrowness of the vocal tract. Any speech sound produced is a combination of the above three states of the speech signal. The primary goal of machine speech recognition is to identify these states of the speech signal.

1.5 Types of Speech Recognition

There are different types of speech recognition systems based on the types of inputs they can take and their ability to recognise.

1. **Isolated Speech Recognition System**: A speech recognition system is called an isolated speech recognition system if each word is preceded and followed by a pause

while speaking [6]. It requires a single utterance at a time. It sets a necessary condition that each utterance has little or no noise on both sides of the sample window.

2. **Continuous Speech Recognition System**: A speech recognition system is called a continuous speech recognition system if the words are spoken in a natural flow [6]. It requires continuous human speech, without silent pauses between words. It is much more difficult to recognise as compared to isolated word recognition.

3. **Speaker-dependent speech recognition system**: A speech recognition system is called a speaker-dependent speech recognition system if it is able to identify the speech of the person for whom it was trained.

4. **Speaker-independent system**: A speech recognition system is called a speaker-independent speech recognition system if it is able to recognise speech from people whose speech was not trained in the system [6]. Such systems can recognise the speech of different people.

5. **Large-vocabulary speech recognition system**: A speech recognition system is called a large-vocabulary speech recognition system if it is trained for roughly 5000 to 60,000 words [6]. Such systems can recognise any of these words while testing.

6. **Small-vocabulary speech recognition system**: A speech recognition system is called a small-vocabulary speech recognition system if it is trained for a lesser number of words.

1.6 Applications of Speech Recognition

Automatic speech recognition by a machine can be very beneficial for day-to-day tasks. Some of the applications of automatic speech recognition are

1. Voice Search [7]
2. Voice Assistants [8]
3. Dictation and Voice Typing [9]
4. Transcription Services [10]
5. Language Learning [11]
6. Accessibility [12]
7. Voice-controlled Systems [13]

8. Medical Transcription [14]
9. Broadcast Captioning [15]
10. Domestic Appliance Control [16]
11. Autonomous Driving [17]

These and many more applications of automatic speech recognition make it an important area of research. Some of the most widely used voice assistants are Google Assistant by Google, Cortana by Microsoft, Siri by Apple, and Alexa by Amazon[18]. A few of these are also available in regional languages.

1.7 Speech Recognition of Indian Languages

India is a country with 22 scheduled languages and 99 unscheduled languages. Indian languages contain more phonemes as compared to other languages. They contain a greater number of retroflex consonants and fricatives. The accents are non-uniform within the same language [19]. Proper articulation is required for correct pronunciation. Hence, recognising a speech in an Indian regional language is a challenging and very important task.

Automatic speech recognition technology is useful in aiding accessibility for people who are deaf or have hearing difficulties. Such systems can transcribe the spoken words of a normal person into sign language in real-time, allowing deaf people to understand the conversation, presentations, or other spoken content. Speech recognition technology can also be used hands-free for those who have limited mobility or some physical disabilities. It is particularly valuable for individuals with disabilities who find it difficult to use regular input devices. With their voice, these individuals can easily give commands to computers, mobile devices, and other smart gadgets. These voice commands may be used to interact with devices or assistive technologies. It can be useful for operating a computer, controlling a domestic appliance, or using any digital interface without the need for conventional input devices like a mouse, keyboard, trackpad, or touch screen. Automatic speech recognition in Indian regional languages is incredibly useful, especially for people who can only speak in their native language. Having speech user interfaces in regional Indian languages would greatly benefit them by making technology more accessible and allowing them to interact with it effortlessly. Our work is on the Gujarati language, which is spoken mainly in the west part of India and has around 62 million speakers [20].

Gujarati Language

Gujarati is a language with 13 vowels and 36 consonants. This language is phonetically different from other Indian languages because it has a retroflex lateral flap like ળ, which is unique in pronunciation. Moreover, the pronunciation of conjunct consonant જ્ઞ is different as compared to other Indian languages. The vowels એ and ઓ each represents two different pronunciations [21]. So a system trained for another language may not work accurately in Gujarati language. Hence, it is essential to make a speech recognition system separately for Gujarati language.

1.8 Literature Survey

There are many methods and algorithms applied for designing an automatic speech recognition system. In the 1950s, early automatic speech recognition systems were based on acoustic-phonetic methods, spectral measurements, and analogue filter banks [5]. These systems were updated in the 1960s using filter bank spectral analyser, time normalisation methods, and dynamic programming methods. The decade of the 1970s added more advancement to it by introducing methods based on pattern recognition, clustering algorithms, and linear predictive coding (LPC) [5]. The 1980s brought a revolutionary change in this field with the replacement of template-based methods by statistical modelling methods like hidden Markov models (HMM) [5]. Artificial neural networks (ANN), expert systems, and wavelets began to be used for automatic speech recognition in the 1990s. In the 21st century, these systems advanced further due to the evolution of machine learning (ML) algorithms and improvisation in ANN due to deep learning (DL). These were all mainly applied to the English language. Our work is for the Indian regional language, Gujarati.

1.8.1 Indian Languages:

Researchers across India are actively working in the research area of automatic speech recognition for different Indian regional languages. A survey of Hindi, Punjabi, Tamil, Assamese, Bengali, Marathi, Oriya, Urdu, Kannada, Telugu, Gujarati, and Bodo languages is summarised in [22]. Similarly, in [23], the progress of speech recognition in Assamese, Bengali, Hindi, Marathi, Oriya, Sinhalese, Urdu, and Punjabi languages is summarised. A recent survey on speech recognition in Indian languages can be found in [24]. A survey of research in speech recognition for Gujarati is summarised in [25].

From these surveys, it can be observed that very little work is done for the Gujarati language. One of the major challenges of automatic speech recognition in Indian languages is the lack of speech corpora [19]. Gujarati is one of the low-resourced languages for speech

recognition. Several researchers are working on generating speech corpora for Gujarati and other Indian languages [26], [27]. Also, a speech corpus in Gujarati language for emotional analysis is generated, as in [28].

1.8.2 Gujarati Language:

For speech recognition in the Gujarati language, many researchers are using different techniques. Approaches like vector quantisation [29], Gaussian mixture models [30], artificial neural networks [31], [32] and support vector machines [33] have been used. In recent works, the hidden Markov models approach has been used [34], [35]. Hidden Markov models together with artificial neural networks are used in [36]. Fast bootstrapping was used in [37]. Some authors have developed the phonetic engine to convert speech sound units into symbols [38]. Syllable boundary detection methods are summarised in [39]. The application "Avaaj Otalo" was developed for people with limited literacy, limited familiarity with technology, and only knowing Gujarati [40]. A speech recogniser of 60 words was prepared for smart-phone operations like calling, sending SMS, and other smart-phone operations using hidden Markov models [41]. Due to the evolution of deep learning techniques, multilingual speech recognition for various Indian languages has been developed. Some researchers have prepared a multilingual speech recogniser that can recognise a speech spoken in various Indian languages [42], [43]. End-to-end systems are proposed [44]. Several researchers have proposed multilingual speech recognition using different DL techniques [45], [46], [47], [48].

Recent studies have focused on advancing Gujarati language-based automatic speech recognition (ASR) systems. Dua et al. [49] reviewed existing Gujarati ASR systems, identifying critical challenges and potential solutions. Integrated feature extraction and hybrid acoustic models were proposed by Dua and Akanksha [50] for improved accuracy. Sharma et al. [51] explored code-switching speech recognition for Gujarati-English, emphasizing words with similar sounds in both languages. Shah and Kavathiya [52] conducted an in-depth study of Gujarati speech corpora for voice recognition, while Bhagat and Dua [53] demonstrated the effectiveness of an improved spell-correction algorithm combined with the DeepSpeech2 model to enhance Gujarati ASR performance. These studies provide a strong foundation for advancing in Gujarati ASR, which this thesis aims to build upon by addressing the under-explored areas of feature extraction and machine learning techniques.

From the literature review, it can be observed that the Gujarati language is a low-resourced language, and a limited amount of work is done for automatic speech recognition in the Gujarati language. Most of the research works for the Gujarati language are

based on feature extraction using the mel-frequency cepstral coefficient. The wavelet-based coefficients are yet to be explored for the Gujarati language for feature extraction. Moreover, machine learning technique like the radial basis function network is also not used for the speech recognition of Gujarati language.

1.9 Objectives

From the literature survey, back in the initial days of our research work, we observed that most of the research works for speech recognition for Gujarati language and the ones that exist are based on feature extraction using the mel-frequency cepstral coefficient (MFCC). MFCC are features based on discrete cosine transforms. Wavelet-based coefficients like mel-frequency discrete wavelet coefficients (MFDWC) are yet to be explored for the Gujarati language for feature extraction. Moreover, machine learning technique like radial basis function (RBF) network is also not used for the speech recognition of Gujarati language. In our work, we have overcome these limitations and compared several methodologies for automatic speech recognition of the low-resourced Gujarati language.

In our work, automatic speech recognition of Gujarati language is done, pertaining to two broad objectives:

- Automatic speech recognition of isolated words, spoken in Gujarati.
- Automatic speech recognition for continuous sentences, spoken in Gujarati.

To achieve the objectives for isolated words, we have used the existing feature extraction technique of mel-frequency cepstral coefficients (MFCC) and compared it with the approach of using feature extraction technique based on mel-frequency discrete wavelet coefficients (MFDWC). In both approaches, the task of classifying features was carried out using the distance-based dynamic time warping (DTW) method and machine learning methods like multilayered perceptron (MLP) and radial basis function networks (RBFN). We also used a method based on the hidden Markov model (HMM) in the case of feature extraction using mel-frequency discrete wavelet coefficients (MFDWC).

For the dataset with continuous sentences, we focused on the approach of mel-frequency discrete wavelet coefficients (MFDWC) for feature extraction. In this case, the classification task was carried out using a multilayered perceptron (MLP) and a hidden Markov model (HMM). In both approaches for the recognition of sentences, we also tried ensemble learning with various models. Finally, we created an interface for speech recognition for Gujarati language using these two approaches. The research works carried out to fulfil out objectives, are summarised in the block diagram shown in the Figure 1.5. Viewing

this diagram as a tree structure, all the leaves of this tree represent our various approaches of speech recognition in Gujarati language. The basic concepts, methods, and techniques related to these approaches are reviewed in the next chapter.

This thesis is organised into a total of seven chapters. In the chapter 2, preliminaries and methodologies required in speech recognition are explained. chapter 3 explains the proposed models for which feature extraction is done using discrete cosine transform based mel-frequency cepstral coefficients and classification is done using dynamic time warping, multilayered perceptrons, and a radial basis function network. The accuracy results of each and their comparisons are also given. Chapter 4 explains the proposed models for which words are extracted from sentences and then features are extracted from words using discrete wavelet transform based mel-frequency discrete wavelet coefficients along with their performance. In the same chapter, the models on the augmented dataset are also explained. Chapter 5 is about ensemble learning models in which various models are combined to get better generalisation. In chapter chapter 6, we discuss our efforts to create a graphical user interface for the speech recognition of Gujarati language. Finally, the thesis ends with chapter 7, which lists concluding remarks and a discussion of future scopes.

1.10 Conclusion

This chapter began with a review of what speech recognition is, along with its importance and basic ideas in section 1.1. We discussed how speech recognition was motivated by human speech perception in section 1.3. The automatic speech recognition task is the machine counterpart of how our ear works. The different types of speech recognition based on recordings and recognition capability are also described briefly in this chapter as section 1.5. There was also a brief discussion on applications of speech recognition in day-to-day tasks in section 1.6. Since our speech recognition work is for Gujarati language, we explained the importance of speech recognition in regional Indian languages and Gujarati language in section 1.7. Moreover, the review of related literature, for the speech recognition tasks, is also summarised in this chapter as section 1.8. Finally, section 1.9 explains objectives and the brief of various approaches.

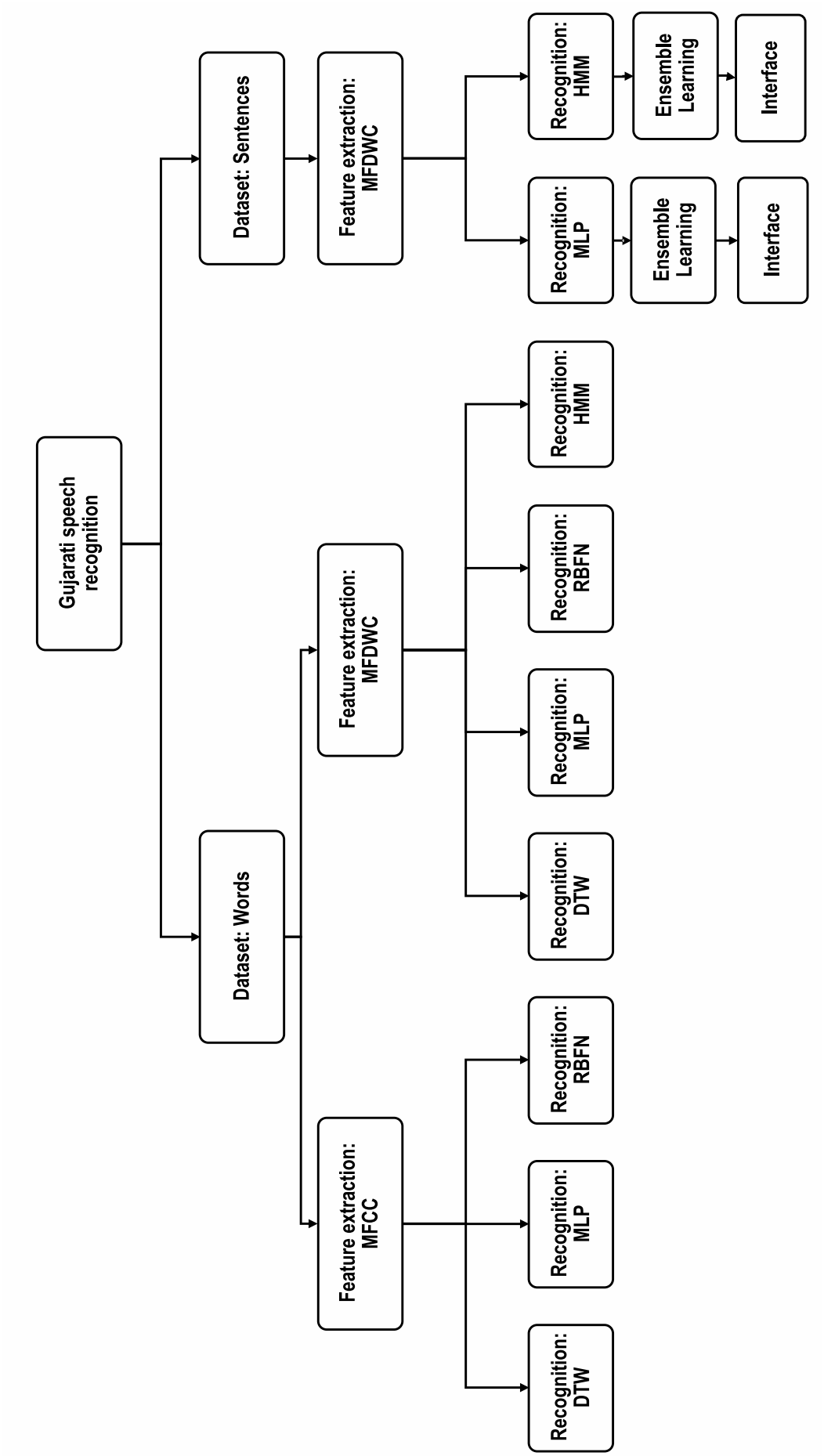

Figure 1.5: Block diagram of a summary of the approaches to our research work