Abstract

In this era of smart gadgets involving artificial intelligence, automatic speech recognition (ASR) is an integral part. ASR is the process of recognising a speech spoken by a human and deriving meaning from it. It has a variety of applications and is also useful for people with disabilities. Speech recognition is a challenging research area because it is an interdisciplinary field of study. The nature of the speech signal adds more challenge to it due to noise and variability in the signal. There is a need for research in the area of speech recognition in Indian regional languages due to the large number of native language speakers in India. Such an ASR system is useful for people who know only the regional language.

This thesis presents our efforts to develop ASR for the Gujarati language. The main objectives of this research are to review, propose, and apply various speech recognition methodologies for the Gujarati language. The process of ASR involves three main phases: pre-processing, feature extraction, and recognition. Our work is a compilation of existing and proposed methods to implement these three phases. These methods fall into the interdisciplinary areas of signal processing, applied mathematics, statistics, probability, machine learning, and computer science.

We have proposed and built various models for the speech recognition of words and sentences spoken in the Gujarati language. We generated datasets for the same by recording the speeches by various speakers, spoken in the Gujarati language. For sentences recorded in the Gujarati language, we proposed a method to extract words automatically from sentences using the signal processing technique of short-term autocorrelation. For feature extraction, we have used techniques like mel-frequency cepstral coefficients and mel-frequency discrete wavelet coefficients. The later one is based on the discrete wavelet transform and gives better results.

For the recognition of speech, various classification techniques like dynamic time warping, multilayered perceptrons, radial basis function networks, and hidden Markov models are implemented in our work. The accuracies of speech recognition for all these techniques are also computed and compared. We have even used the data augmentation technique

in our work to generate more samples of data for better performance by the models. The use of ensemble learning is also explored to improve recognition accuracy. Further, the final outcome is the graphical user interface for the speech recognition of Gujarati sentences. The interface has the capacity to record the speech, play it, and then recognise a sentence spoken in the Gujarati language using the trained model based on classification techniques for a small vocabulary.

The thesis has chapters arranged in the following manner: Chapter 1 contains a description of the processes of speech recognition along with the brief structure of Indian languages. It also contains sections on the literature survey and objectives. Chapter 2 describes the preliminaries and methodologies required to build a speech recognition model.

In chapter 3, we propose models for speech recognition of Gujarati words, particularly digits. In this work, the feature extraction technique, the mel-frequency cepstral coefficient, is integrated with classification techniques like dynamic time warping, multilayered perceptrons, and radial basis function networks. In chapter 4, we propose models for speech recognition of Gujarati words and sentences based on feature extraction using mel-frequency discrete wavelet coefficients. For the recognition of sentences, before feature extraction, we integrate the short-term autocorrelation technique for word extraction. Here, we use dynamic time warping, multilayered perceptrons, a radial basis function network, and hidden Markov models for the classification of Gujarati words and the multilayered perceptron and hidden Markov model for the recognition of sentences. In this chapter, data augmentation techniques are also used to expand the dataset and improve the performance of models.

In chapter 5, we have combined several speech recognition models using the approach of ensemble learning to improve the recognition accuracy. Chapter 6 explains the graphical user interface that we have created for the speech recognition of Gujarati sentences. The thesis ends with a chapter 7 having conclusions and future scope.