Chapter-1

Introduction

Human beings need speech for communication. The information is contained in the speech which is needed for the listener and speaker. It also contains message including speaker's characteristics like emotion, information regarding the language and his physiological characteristics. The environmental information [1] in which the communication was performed between the sender and receiver is also contained in the speech. The speech signal contains large amount of information which is complex and in a coded form, but due to the intelligence of humans, these can be decoded easily. The automation in the field of human machine interaction is developed by many researchers for understanding the production of speech and extraction of information contained in the speech. The various applications [2] of this technology are audio indexing and retrieval, control using voice command, field of dictation etc. The emotion of the speaker can be found out from his speech based on his accent apart from the content and his nativity. The speech quality deteriorates with noise which is an unwanted signal and is irrelevant.

1.1. Importance of speech/ speaker recognition

Speech/Speaker recognition is a growing technology in different areas for its various applications. In order to provide a comfortable and natural form of communication between speaker and personal computer, speaker recognition becomes more important technological development. Using the voice only, the amount of

typing can be reduced, hands can be free and also movement away from the computer can be possible along with security. Hence speech recognition is a highly motivated research area

The main purpose of speech processing [3] is designing a machine that mimics human behavior, particularly the capability of speaking naturally and responding properly to spoken language. Speaker recognition has a history back some four decades and uses the acoustic features of speech, that have been found to differ between individuals. These acoustic patterns reflect both anatomy (e.g., size and shape of the throat and mouth) and learned behavioral patterns (e.g., voice pitch, speaking style).

1.2. Basics of Speech production

Speech is an important time varying means of communication for human beings. It is produced by exciting the vocal tract system which time is varying. The vocal chords vibrates when the air is expelled from the lungs. The quasi-periodic pulses are produced due to these vibrations in glottis and subsequently there is air flow into supra-glottal region, which is above the glottis. The pharynx, nasal cavity and the oral cavity areas together constitute as supra glottal region. The larynx is below the pharynx area and uvula (behind the throat a blob of flush that hangs) is above the pharynx as shown in the fig 1.1. The various articulators (like lips, teeth, tongue etc.) are modified due to the quasi-periodic pulses, which are generated when the air flows through the oral cavity and pharynx. Further, lowering velum also modifies the airflow when it passes through naso-pharynx and the nostrils. These modifications of the pulses and the air flow in the complete system causes the radiation of sound waves at the nostrils and lips.

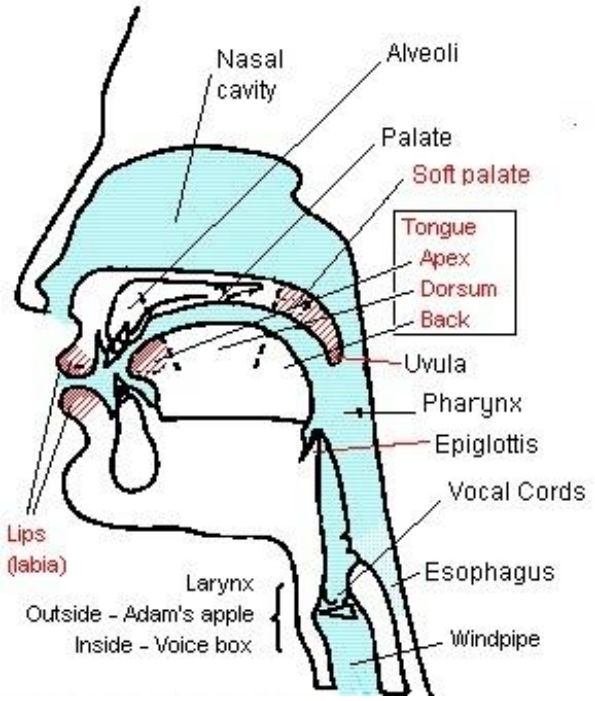

Fig 1.1: Human speech production system

When the excitation is from vibrating vocal folds, production of voiced sounds is achieved. Turbulence is observed, when the air is forced to flow through a narrow constriction and this results in unvoiced sounds. The air particles random movement causes the turbulence in comparison to the quasi-periodic air flow. In the case of unvoiced sounds, the source of the noise in the vocal tract might be located at different places, with respect to place of articulation of the sound. The position of the noise generator affects the dimensions of the section in the vocal tract. The noise will be modified based on the vocal tract sections of both the front and backside of the constriction. In some frequency bands, it can be observed that there is sudden decrease in the noise energy amplitude. These are found for fricative sounds. For the 'stop' consonants, a very short noise burst is obtained due to rapid release. The vocal tract filter modifies 'stop' consonant based on

articulation place of the sound. Due to the articulator's flexibility, infinite modifications are possible in the production of speech.

The whole frequency spectrum with many speech components can be obtained from the speech which is recorded using a simple microphone. The simple micro phone is preferred, due to the signal's high perceptual quality and intelligibility.

1.3 Development in speech recognition

The researchers attempted to work in the area of speech processing on acoustic-phonetics way back in the early 1950's. The research [4] of digits which are isolated for a single speaker was attempted in the year 1952. The complete speech was segmented and the speech recognition algorithms were developed during the initial phase of Automatic speech recognition (ASR) systems. These identified segments helped for recognizing any individual's speech. The development in electronics during the initial years of 1970, large amount of flexibility to use processors increased the speed of development in the area of research in speaker recognition. An algorithm for connected word recognition was proposed by Sakoe and Chiba [5] in 1979. Similar to the Dynamic programming algorithm, a new technique was proposed by Rabiner and Myers [6], which was termed as Dynamic Time Warping (DTW) in 1981. This proposed system was complex but, however it was efficient and flexible. Hidden Markove Model (HMM) [7] was applied for research in ASR in the beginning of 1980's. The research in this area is still a open area and lot of development is required in future.

1.4. Accent recognition

Accents of a given language are formed due to speaking style differences of a particular language and geographical an ethnic differences of the speakers. A difference in acoustic space spanned by phonemes for native speakers and non-native speakers can be observed. The intonation, duration, rhythm, voiced stop release time play vital role in recognizing the accent. The identification accent will increase the performance of the speech recognition systems as it limits the searching to the recognized accent. Accent recognition system improves Human Computer Interaction. In IVRS systems, the machine can interact with the user in their own accent after recognizing the accent.

There are different levels of information in speech with respect to accent. In the **segmental level**, various units of sound are produced based on the accent of the speakers produced by the vocal tract. The Mel Cepstral Coefficients decides the spectral envelope for the vocal tract. The vocal folds of open/ close phase duration contain the information about the accent. The energy features, pitch dynamics, duration, dialect are some of the information contained in a speech signal.

In the sub segmental level, information with respect to dialect might be available in the form of glottal pulse with time periods pertaining to the open and close phases for vocal folds.

The features belonging to segmental and supra segmental levels can be used for recognition of accent of any language. In general 20-30ms time period speech segment is used for extracting features belonging to segmental level. These are also called as

spectral features since, they are contained in the frequency spectrum which is found in the speech segment. The prosodic features have a time period of more than 100ms and are also called supra-segmental features. The duration of speech segment is less than 3ms for sub-segmental features.

There are reports in the literature survey, about the existence of a few studies with respect to automatic identification of dialect in Japanese and English [1, 5]. However, for Indian languages dialect analysis is at nascent stage. Further, it is found in the literature survey that very less amount of work is done using the features extracted from the speech for dialects. Hence, due to this motivation, the present proposed work deals with feature extraction from the speech and analysis of dialects for Telugu speech.

There are many accents of Telugu language which are spoken in different regions of Telugu states i.e. Andhra Pradesh and Telangana. However, broadly there are three main accents namely Coastal Andhra, Rayalaseema and Telangana which are considered in this proposed research work.

1.5. Emotion recognition

The signals that are derived from the speech samples exhibit the identity of the speaker and further investigation these speech samples help to interpret the behavioral state of mind of the individual. The speech signal that is generated from the uttered speech consists of both the out word physical expression and the inherent emotional feelings. In general every individual exhibit the

same inherent feeling for a specific or similar expression. The prosody of the speech signal is associated with every speech emotion. This prosody depends upon several factors which include the living condition, the language rules residential locating of the individual community culture etc. In general emotional recognition is a methodology of synthesizing the individual's inherent behavior by considering a fragment of speech sample uttered by the individual. Also it can be interpreted as the science of reading the individual mind. Every individual exhibit a specific emotion while reacting to specific situations. One of the greatest challenges is identifying the emotional state of mind of an individual while reacting to a real world situation. The effectiveness to recognize the emotion state of a person helps to promote the social interaction and the interpersonal interaction. In ordered to identify the emotional state of mind of an individual the speech signal is considered and its symbolic representation helps to identify the emotional coefficient.

1.6. Challenges for speech recognition with reference to Indian languages

Speech recognition becomes a very important area of research since people need to communicate with the computer even in Indian languages for many applications. In the area of mobile technology, it is extensively used specially when speech is to be converted to text or vice-versa. In this case it has advantage of not using the keyboard and hands-free communication. This type of need arises when a person is driving his vehicle (four wheeler) and needs hands-free communication in a mobile phone. The technology development in modern days is mostly digital and hence, this helps for the area of speech recognition. Even for the

people who can't read or write a language, but can speak and understand a particular language, the speech recognition technologies help them to understand the information.

India is multi lingual country and has about 1652 dialects from native languages [8]. The official language in India is Hindi written in Devanagari script. However the other languages include Assamese, Tamil, Malayalam, Gujarati, Telugu, Oriya, Urdu, Bengali, Sanskrit, Kashmiri, Sindhi, Punjab, Konkani, Marathi, Manipuri, Kannada, Nepali, Santhali, Bodo, Maithili and Dogri. Hence, the speech recognition system and its synthesis become very important for all the Indian languages mentioned above [5].

1.7. Importance of Speech databases

To develop algorithms and perform research in the area of speech processing, the first step is to generate a database, as per the application. This database could be used for both training and testing the various speech models. In the early phases, research was more concentrated on acoustic-phonetics and the basic idea of research was planned. In the next phase speech databases were developed for connected word recognition, text dependent and text independent speech signals. The various areas of research using speech databases are speaker identification, verification, emotion, gender, regional and social dialects. There are standard databases for English, German and Spanish languages, which are available in the websites. Some of these databases are TIMIT database, TSP speech database and BERLIN database which provide platform to improve the recognition accuracy of the proposed speech algorithms. Research in the area of speech and speaker recognition is at nascent stage for most of the Indian languages. Hence a

separate database for Indian languages, based on the requirement should be developed in the laboratory environment [5].

1.8. Motivation for the Research problem

There is a huge amount of development in the field of digital electronics and computers which help for developing and implementing speech processing algorithms for better human machine interaction. This area of research is called machine learning. The field of speech recognition involves extracting the information in the speech for any language. This is achieved by segmenting the speech sentence into word or group of words, known as diarization. In the literature surveyed, there are systems available with good recognition accuracy even for databases of 30-60 persons [9].

Speech is also considered as a reliable and an important biometric feature for researchers in the area of recognition/ identification. The data acquisition systems for speech is easy and very simple compared to other biometric systems like scanning a finger print or a retina of any human being. High quality mobile phone or a telephone network can be utilized even when people are not cooperating, provided the recording device is of high quality.

Dialect/ accent refers to different ways of pronouncing/ speaking a language within a community. There are mainly three different types of accents in the states of Andhra Pradesh and Telangana. Due to the bifurcation of state, recognition of the speaker to a particular zone/area becomes highly essential which can be known using speaker's accents/ dialects. This would help in

Automatic Speech Recognition (ASR) systems by adapting the acoustic or language models appropriately. This helps in understanding and finding out the nativity of the speaker to classify him for a particular region in the state. Also further after identification of the speaker to a particular region, the speech can be more accurately recognized by the system. There are many challenges for the speech and speaker recognition systems, which motivated for the present work. Every individual has his own speaking style, which depend on dialect, accent and also his socio economic back ground. These individual differences create difficulties in modeling large scale speaker independent systems to process input from any variant of given language. This work focuses on automatically identifying the dialect or accent of the speaker when a sample speech is given. Since there is no standard database for Telugu speeches [5], laboratory environment is used to develop the training set and also the testing set. This further increases to understand the content of the speech and hence speech recognition can be simplified. In this work apart from accent based speech and speaker recognition, emotions are also considered for identifying the speech and speaker. Due to all these challenges in the proposed work, it is felt that research in Telugu speech is highly essential.

1.9. Objectives of the research proposal

In this research work accents and emotions are used for Telugu speech samples to identify/classify the speech and speaker. The following are the objectives of the proposed research.

1. To design and develop algorithms for accent based recognition system for Telugu speech.

In order to achieve this objective the following tasks are to be carried out.

(i) Develop training and testing database containing the accents of Coastal Andhra, Rayalaseema, and Telangana regions of Telugu speech.

(ii) Extract various features and select the most important features for speech identification/classification.

(iii) Design various algorithms that uses the above extracted features for classification/identification.

(iv) Compare the results obtained in the proposed algorithms with the published results.

2. To design and develop algorithms for emotion based recognition system of Telugu speech samples.

(i) Develop training and testing database containing the emotions like Happy, boredom and Neutral of Telugu speech.

(ii) Extract various important features and develop algorithms for identifying Telugu speech/speaker.

(iii) Design classification techniques for achieving high recognition accuracy.

(iv) Compare the results achieved of the proposed algorithms with the published results.

1.10. Organization of thesis

This thesis consists of six chapters, which include introduction and conclusions. Introduction to the research problem, i.e. accent recognition and emotion recognition has been presented in chapter-1. Chapter 2 consists of summary of reported literature collected for various topics like survey on Speech databases, Text dependent speaker recognition, Speaker emotion recognition, Speaker accent recognition, Text independent speaker recognition, Language identification, Speech to Text conversion.

Chapter-3 focuses on the development of algorithms in the area of accent recognition for Telugu speeches using prosodic features. Further, the data acquisition, feature extraction and classification details are discussed thoroughly. In chapter-4, spectral features are used to train the GMM (Gaussian Mixture Model) model. The results obtained using these MFCC (Mel Frequency Cepstral Coefficients) features and GMM are discussed in detail. In this same chapter, three additional features namely pitch chroma, tonal power ratio, spectral flux are considered along with MFCC features. The DBN (Deep Belief Network) concept is used to classify the various accents of Telugu speech.

Chapter-5 deals with the emotion recognition algorithm including the data acquisition, proposed methodology and results. The comparison between the proposed and published result is also discussed in detail. The overall conclusions of the proposed work and also the future scope are presented in Chapter-6.

1.11 Conclusions

The basic concepts and production of speech is discussed in this chapter. The importance, challenges and the motivation for the research in speech processing area are also presented in this chapter. The need for database generation and introductory concepts about the accent and emotion recognition are thoroughly discussed. This chapter also has included the various applications of speech processing. The objectives and the various tasks involved to achieve these objectives are also discussed in this chapter. The organization of the thesis is included, explaining the contents of each chapter of this research work.