Hearing and Speech processing

Introduction

Speech is primary mode of communication among human beings and also the most natural and efficient way of exchanging information among humans. Speech can be defined as the expression of or the ability to express thoughts and feelings by articulate sounds. No two individuals' sound identical because their vocal tract shapes, larynx sizes, and other parts of voice production organs are different. In addition to these physical differences, each speaker has his or her own characteristic manner of speaking, including the use of a particular accent, rhythm, intonation style, pronunciation pattern, choice of vocabulary and so on.

Speech signal processing refers to the acquisition, manipulation, storage, transfer and output of vocal utterances by a computer and is the process of converting spoken input to text. An important goal of speech processing is to develop techniques for the systems to accept speech as input. Recognition, synthesis and compression of human speech form key areas in speech processing. Given a speech signal, there are two kinds of information that may be extracted from it. On one hand, there is the linguistic information about what is being said (speech recognition), and on the other hand, there is speaker specific information (speaker recognition).

Speech recognition (or Automatic Speech Recognition) focuses on capturing the human voice as a digital sound wave and converting it into text form and is sometimes referred to as *speech-to-text*. Synthesis is the reverse process of speech recognition which transforms text into speech and is also referred to as *text-to-speech*. Advances in this area improve the computer's usability for the visually impaired.

Speaker recognition (or Automatic Speaker Recognition) on the other hand is concerned with recognizing the speaker. Speaker recognition process is classified as *open set* if the test utterances spoken by the speakers do not form part of the training data. Similarly, if test utterances belong to the speakers in the training data, the set of speakers is called a *closed set*. A *text dependent system* relies on the restriction that the text in training and test utterances is identical. On the other hand, a *text independent speaker recognition system* does not require the content of training and test utterances to be the same.

Speech compression is important in the telecommunications area for increasing the amount of information which can be transferred, stored, or heard, for a given set of time and space constraints. Automotive speech recognition, court reporting (Real-time Speech Writing), hands-free-computing, home automation, interactive voice response, robotics speech-to-text reporter are some of the applications of speech recognition. Voice dialing, banking by telephone, telephone shopping, database access services, voice mail, security control for confidential information areas, remote access to computers, forensics (as much of information is exchanged between two parties in telephone conversations, including between criminals) are some of the applications of speaker recognition.

Review

Early Neurolinguistics Models

Wernicke-Lichtheim-Geschwind model is primarily based on research conducted on brain-damaged individuals who were reported to possess a variety of language related disorders. In this model, words are perceived via a specialized word reception center (Wernicke's area) that is located in the left temporoparietal junction. This region then projects to a word production center (Broca's area) that is located in the left inferior frontal gyrus. Because almost all language input was thought to funnel via Wernicke's area and all language output to funnel via Broca's area, it became extremely difficult to identify the basic properties of each region. This lack of clear definition for the contribution of Wernicke's and Broca's regions to human language rendered it extremely difficult to identify their homologues in other primates. With the advent of the fMRI and its application for lesion mappings, however, it was shown that this model is based on incorrect correlations between symptoms and lesions. The refutation of such an influential and dominant model opened the door to new models of language processing in the brain.

The auditory ventral stream (AVS) connects the auditory cortex with the middle temporal gyrus and temporal pole, which in turn connects with the inferior frontal gyrus. This pathway is responsible for sound recognition, and is accordingly known as the auditory 'what' pathway. Accumulative converging evidence indicates that the AVS is involved in recognizing auditory objects. At the level of the primary auditory cortex, recordings from monkeys showed higher percentage of neurons selective for learned melodic sequences in area R than area A1, and a study in humans demonstrated more selectivity for heard syllables in the anterior Heschl's gyrus (area hR) than posterior Heschl's gyrus (area hA1).

Early Speaker Recognition Systems

In the early 1960’s, law enforcement agencies approached the Bell Laboratories about the possibility of identifying callers who had made verbal bomb threats over the telephone. This resulted in the birth of the concept of speaker recognition with the physicist Lawrence Kersta publishing in Nature the research in an article entitled, “Voiceprint Identification”.

Many other parallel developments in the field contributed to the field of autonomous speech recognition. These developments covered a broad range of disciplines. For instance, Gunnar Fant produced a physiological model of human speech production that became the basis for understanding how to analyze speech. Research into the physiological aspects of voice led future researchers to represent voice as a linear source-filter type model. Understanding voice using such a model allowed for many advances in discovering identifiable characteristics in an individual’s voice.

During the same decade, several other investigations into automatic speaker recognition also bore results. Pruzansky (Bell Laboratories) investigated, with limited success, a speaker recognition system utilizing spectral pattern matching. The first successfully implemented autonomous speaker recognition system was developed by a team led by George Doddington at Texas Instruments in 1977.

Work Done

Speech processing is the study of speech signals and the processing methods of signals. Speaker recognition, a subset of speech processing, is concerned with determining the identity of a person based on his/her utterances stored in the database.

In this project, a simple text-independent speaker recognition system has been developed using MATLAB. The training and test utterances need not be same in the data. The system accepts speech utterances for a speaker as input. The goal of the system is to determine which one of a group of known voices best matches the input utterance.

The process involves first extracting the MFCC features of a group of known speakers. These features are used in the training phase to build a model of the known speakers. In the recognition phase, given a test speaker, the system extracts the MFCC features (as in training phase), and passes these features to the speaker model which determines a unique ID corresponding to the test speaker. Experiments show a high accuracy even in the presence of some noise.