2. SPEAKER RECOGNITION CONCEPTS

2.1 Overview

Development of speaker identification systems began as early as the 1960s with exploration into voiceprint analysis, where characteristics of an individual's voice were thought to be able to characterize the uniqueness of an individual much like a fingerprint. The early systems had many flaws and research ensued to derive a more reliable method of predicting the correlation between two sets of speech utterances. Speaker identification research continues today under the realm of the field of digital signal processing where many advances have taken place in recent years.

The problem of speaker identification is one that is rooted in the study of the speech signal. A very interesting problem is the analysis of the speech signal and therein what characteristics make it unique among other signals and what makes one speech signal different from another.

Digital signal processing (DSP) is the processing of signals by digital means [27][5]. In many cases, the signal is initially in the form of an analog i.e.., electrical voltage or current and through one of the processing techniques a discrete or digital output analogous to the analog signal is produced. Signals commonly need to be processed in a variety of ways. Today, the filtering of signals to improve signal quality or to extract important information is done by digital signal processors using DSP techniques rather than by analog electronics.

Although the mathematical theory underlying DSP techniques, such as the Fast Fourier Transform (FFT), digital filter design and signal compression can be fairly complex, the numerical operations required to implement these techniques are in fact very simple . Although FFT is not very optimal in many filtering applications it is a very commonly used technique for analyzing and filtering digital signals.

The digital signal processor is a programmable microprocessor device with its own native instruction codes that is capable of carrying out millions of floating point operations per second. By coding the various digital signal processes theoretically, a digital signal processor can be replicated using a data-manipulation and development Environment such as Matlab.

Speech processing is a diverse field with many applications. Figure 2.1 shows a few of these areas and how speaker recognition relates to the rest of the field.

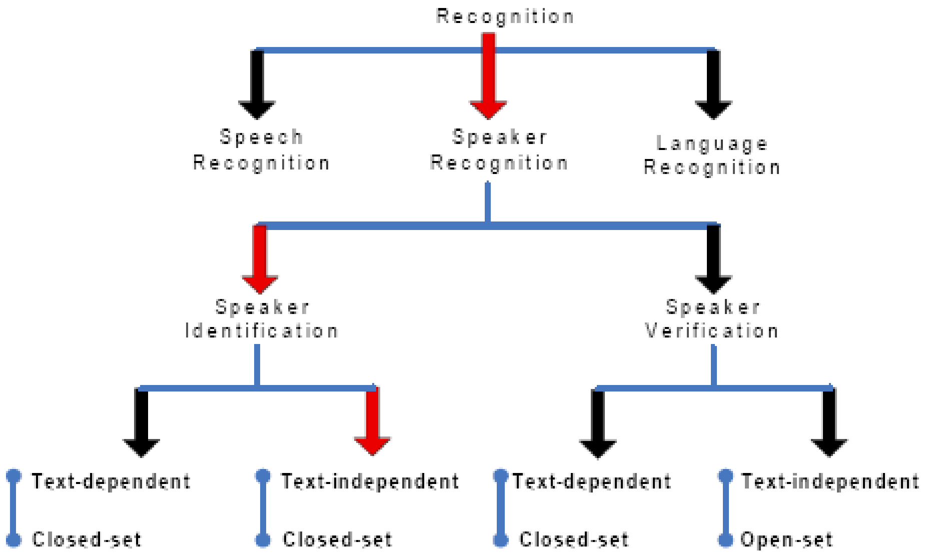

Figure 2.1 Speech Processing [19]

Speaker recognition is the process of automatically recognizing who is speaking based on unique characteristics contained in speech waves. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers. All speaker recognition systems at the highest level contain two modules, feature extraction and feature matching. Feature extraction is the process of extracting unique information from voice data that can later be used to identify the speaker. Feature matching is the actual procedures of identifying the speaker by comparing the extracted voice data with a database of known speakers and based on this a suitable decision is made.

A typical Speaker Verification setup is shown in Figure 2.2 [1]. The claimant, who has previously enrolled in the system, presents an encrypted smart card containing his identification information. He then attempts to be authenticated by speaking a prompted phrase(s) into the microphone. There is generally a tradeoff between recognition accuracy and the test-session duration of speech. In addition to his voice, ambient room noise and delayed versions of his voice enter the microphone via reflective acoustic surfaces. Prior to a verification session, users must enroll in the system (typically under supervised conditions). During this enrollment, voice models are generated and stored (possibly on a smart card) for use in later verification sessions. There is also generally a tradeoff between recognition accuracy and the enrollment-session duration of speech and the number of enrollment sessions.

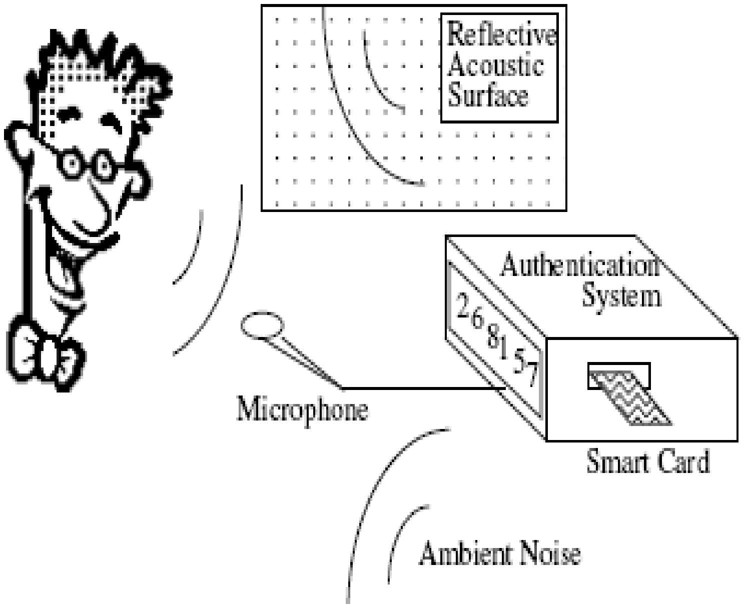

Figure 2.2 Typical speaker verification setup. [21]

Many factors can contribute to verification and identification errors. Table 2.1 lists some of the human and environmental factors that contribute to these errors. These factors are generally outside the scope of algorithms or are better corrected by means other than algorithms (e.g., better microphones). However, these factors are important because, no matter how good a speaker recognition algorithm is, human error (e.g., misreading or misspeaking) ultimately limits its performance.

Table 2.1 Sources of Verification Error [1]

Misspoken or misread prompted phrases
Extreme emotional states (e.g., stress or duress)
Time varying (intra- or intersession) microphone placement
Poor or inconsistent room acoustics (e.g., multipath and noise)
Channel mismatch (e.g., using different microphones for enrollment and verification)
Sickness (e.g., head colds can alter the vocal tract)
Aging (the vocal tract can drift away from models with age)

Figure 2.3 shows a conceptual presentation of a speaker identification system where the input speech initially goes through a feature extraction section where a model of the speech signal is formed. This model is then compared to a reference model database which stores models of all the speakers. The required speaker is found when there is a match between a database model and the speaker model.

There are many techniques used to parametrically represent a voice signal for speaker recognition tasks. These techniques include Linear Prediction Coding (LPC), Auditory Spectrum-Based Speech Feature (ASSF), and the Mel- Frequency Cepstrum Coefficients (MFCC). The MFCC technique was used in this thesis. The MFCC technique is based on the known variation of the human ear’s critical bandwidth frequencies with filters that are spaced linearly at low frequencies and logarithmically at high frequencies to capture the important characteristics of speech.

Speaker recognition has a history dating 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). Speaker verification has earned speaker recognition its classification as a "behavioral biometric."

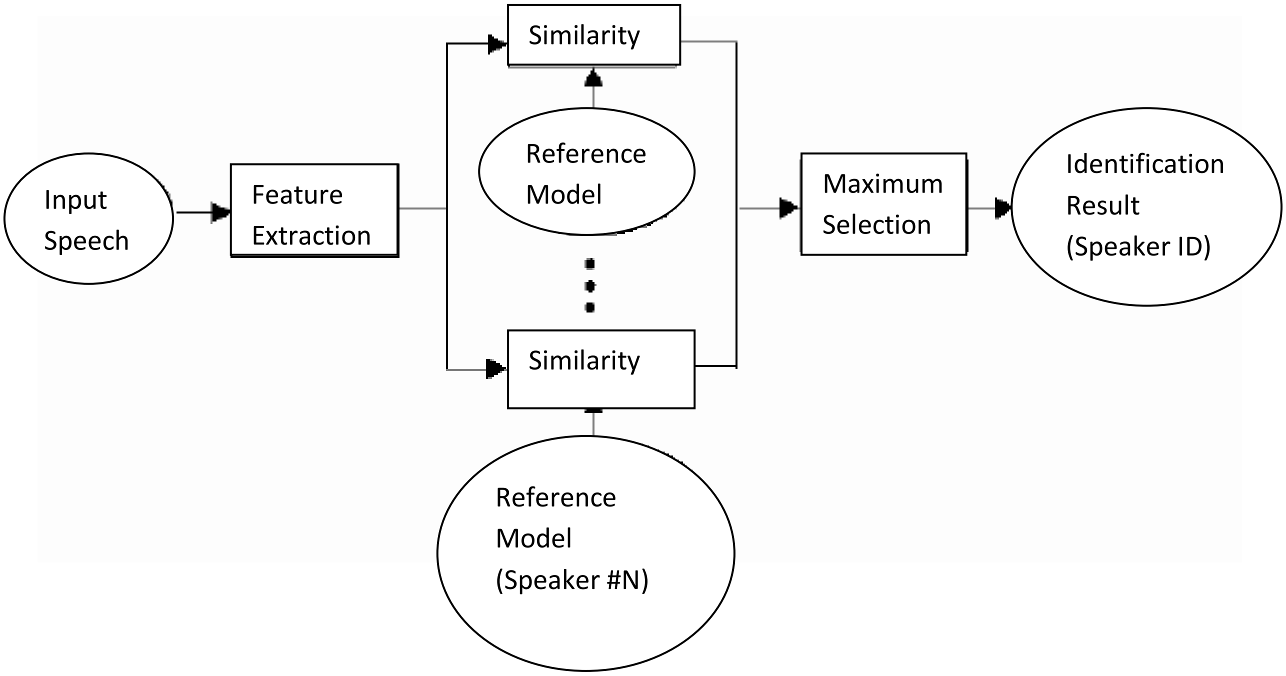

Figure 2.3 Conceptual presentation of speaker identification. [1]

2.1.1 Problem statement

Speaker recognition is usually a general name referring to two different subtasks: speaker identification (SI) and speaker verification (SV). The research in this thesis primarily concentrates on the speaker identification task. The aim in SI is to recognize the unknown speaker from a set of known speakers (*closed-set* SI). Referring to the Figure 2.4 below, one can see that a speaker recognition system is composed of the following modules

Front-end processing : The "signal processing" part, which converts the sampled speech signal into set of *feature vectors*, which characterize the properties of speech that can separate different speakers. Front-end processing is performed both in *training*- and *recognition* phases.

Speaker modeling: This part performs a reduction of feature data by modelling the distributions of the feature vectors.

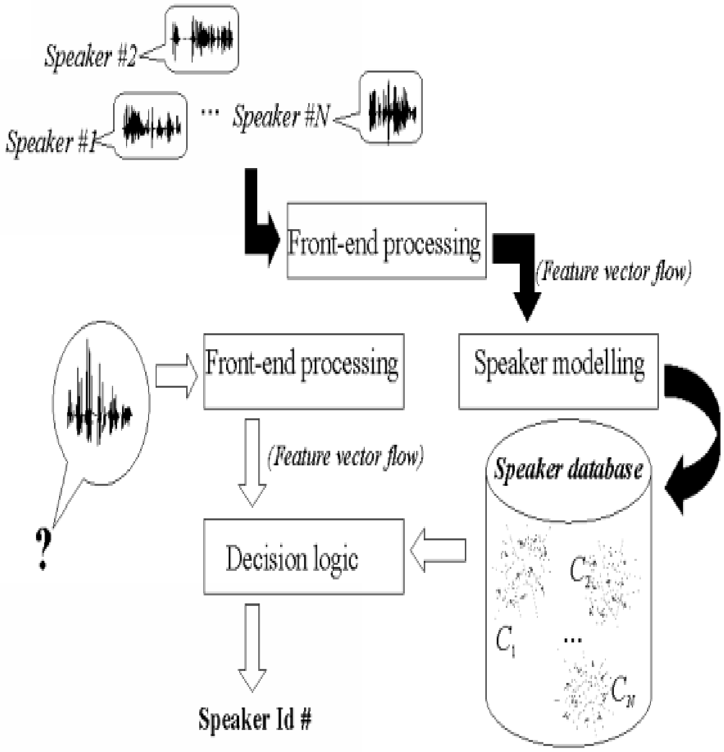

Figure 2.4 Schematic diagram of the closed-set speaker identification system. [21]

Speaker database: The speaker models are stored here.

Decision logic: Makes the final decision about the identity of the speaker by comparing unknown feature vectors to all models in the database and selecting the best matching model.

The first three modules constitute the feature extraction part of speaker identification, while the last module constitutes the feature matching part.

2.2 Biometrics

Biometrics refers to the automatic identification of a person based on his/her physiological or behavioral characteristics. This method of identification is preferred over traditional methods involving passwords for various reasons.

(i) The person to be indentified is required to be physically present at the point-of-identification.
(ii) Identification based on biometric techniques obviates the need to remember a password.
(iii) Proxy methods of impersonation will fail.

By replacing passwords, biometric techniques can potentially prevent unauthorized access to or fraudulent use of ATM's, cellular phones, desktop PCs and computer networks. Moreover biometric systems cannot be rendered to forging.

Various types of biometric systems are being used for real-time identification; the most popular are based on face recognition and fingerprint matching. Other biometric systems utilize iris and retinal scan, speech, facial thermo grams, and hand geometry. An important issue in designing a practical system is to determine how an individual is identified. Depending on the context, a biometric system can be either a verification (authentication) system or an identification system.

Biometric systems automatically recognize a person using distinguishing traits (a narrow definition). Speaker recognition is a performance biometric; i.e., you perform a task to be recognized. Your voice, like other biometrics, cannot be forgotten or misplaced, unlike knowledge-based (e.g., password) or possession-based (e.g., key) access control methods. Speaker-recognition systems can be made somewhat robust against noise and channel variations, ordinary human changes (e.g., time-of day voice changes and minor head colds), and mimicry by humans and tape recorders.

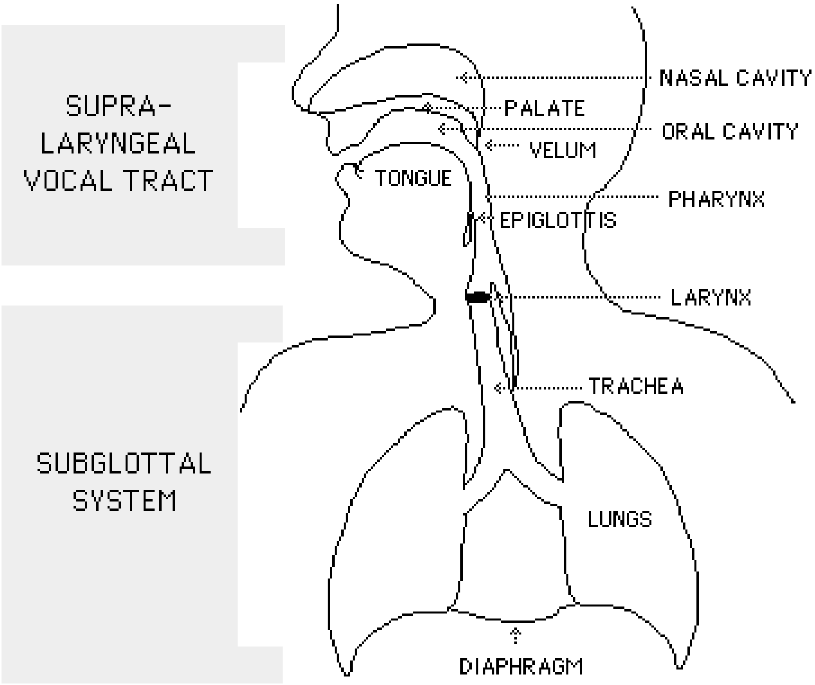

Figure 2.5 Human Speech Production System. [26]

The known biometric methods are shown below:

- DNA Structure Analysis
- Ear Footprint Recognition
- Face Recognition
- Signature Approve
- Recognize from Walking Manner
- Hand and Finger Geometry Recognition
- Iris Recognition
- Retina Recognition
- Smell Recognition
- Fingerprint Recognition

- Speech Recognition
- Speaker Recognition

In general a person can be authenticated in three different ways

- Something the person has, e.g. a key or a credit card, signature.
- Something the person knows, e.g. a PIN number or a password.
- Something the person is, e.g., fingerprints, voice, facial feature.

2.3 Relevant Studies

The research on speaker recognition systems is not new. Many people have contributed to the development of speaker recognition systems. This section gives a brief survey of the relevant literature on speaker recognition systems.

Hasan et al. (2004) describe the development of a speaker identification system using Mel Frequency Cepstral Coefficients (MFCC). The authors used both a triangular and a hamming window to examine the identification rate and report that the results with the hamming window were more satisfactory. The authors also suggest that in order to obtain satisfactory results, the number of centroids had to be increased as the number of speakers increased.

Muda et al. (2010) describe a voice recognition system using the MFCC and Dynamic Time Warping (DTW) techniques. The authors report that both techniques could be used effectively for voice recognition purposes.

Ellis (2001) give details of a project using the MATLAB to design a speaker recognition system. The system is based on pitch analysis for the recognition of a speaker. MATLAB code is presented detailing the steps necessary for pitch analysis in the presence of noise.

An automatic speaker recognition system based on MFCC is presented by Velisavljevic (2003). The system is in the form of a mini-project and is based on MATLAB. Full MATLAB code is given in the project as function files.

Xiong (2006) describes the design of a speaker identification system based on the MFCC technique, using 12 coefficients. Four different version of the system was implemented with different feature extraction methods. The author reported that the four different training methods showed similar performances in accuracy. The implementation consisted of data collection, silence removal, signal emphasis, MFCC calculation, signal shifting, training, and evaluation.

2.4 Summary

This chapter describes, the concepts of speaker recognition, the speaker processing groups (speaker identification, speaker verification) and the methods of the speaker recognition systems (text dependent, text-independent) an automatic speaker identification scheme is to identity or verify a person, by identifying his/her voice.