1. Introduction

The speech signal conveys levels of information. Primarily, it conveys the message being spoken. However, the signal also conveys information about the identity of the speaker, his physiologic and psychological states. The goal of any speaker recognition system is to identify the person uttering the given speech signal.

Humans are fairly adept at speaker recognition. They can easily determine the identity of familiar speakers on the telephone after listening to only a short segment of speech. However, they are less effective in recognizing the voices of unfamiliar speakers. Nevertheless, they can be "trained" to do this by becoming familiar with a given speaker's voice. Designers of speaker recognition systems try to roughly duplicate the processes humans use to recognize speakers. Such systems must be trained to know how a given person "sounds," and the more training data there is, the better are the results.

Automatic Speaker Recognition (ASR) is the process by which a speaker can be automatically recognized by using speaker-specific information included in speech waves. ASR systems can be divided into *speaker verification* and *speaker identification*. Automatic Speaker Verification (ASV) is the use of a machine to verify a person's claimed identity from his voice. In Automatic Speaker Identification (ASI), there is no a priori identity claim. The system has a list of known speakers and it has to classify the unknown speaker to one of the known speakers. Automatic speaker identification systems can further be divided into opened set ASI and closed set ASI. In the closed set system, it is given that the unknown speaker is indeed one of the known list of speakers, while in the opened set system the unknown speaker may not belong to the list, and hence a rejection policy must be included.

Speaker recognition systems can also be divided into *text-dependent* and *text-independent* systems. The former requires the speaker to provide utterances of key words or sentences that are identical for both training and recognition whereas the latter does not rely on a specific text being spoken.

Automatic speaker recognition is based on a biometric measure. Biometric measure is used to differentiate techniques that base identification on certain intrinsic characteristics of the person (such as voice, fingerprints, retinal patterns, or genetic structure) from those that use artifacts for identification (such as keys, badges, magnetic cards, or memorized passwords). This distinction confers upon biometric techniques the implication of greater identification reliability, perhaps even infallibility, because the intrinsic biometrics are presumed to be more reliable than artifacts, perhaps even unique. Convenience is another benefit that accrues to a biometric system, since biometric attributes cannot be lost or forgotten and thus need not be remembered [Doddington, 1985].

Speaker recognition is perhaps the most natural method to solve problems related to unauthorized use and multilevel access control, especially over telephone lines. Furthermore, speaker verification systems can be made resilient to attack from mimicry by humans and tape recorders. However, unlike other intrinsic characteristics, speech is not static. The voice is highly sensitive to the speaker's physiologic and psychological states. Environment conditions (heat, coldness), psychological stress, hoarseness, etc. change person's speech characteristics, and make the recognition task more difficult. Speaker recognition systems have to cope with this problem.

Automatic speaker recognition can be used in many areas: commercial, forensic, military, and medical. In commercial applications, the user is most often cooperative. These

applications include access control by voice, in various services: voice dialing, banking transactions over a telephone network, telephone shopping, database access services, information and reservation services, voice mail, security control for confidential information area, and remote access to computers. On the other hand, the users of forensic and military applications are most often non-cooperative, which further complicates the recognition process. Medical applications use the speakers' speech features to extract information on the physiological states for diagnosis purposes.

Figure 1.1 shows a basic speaker recognition system. It involves training, testing (identification/verification) and memory. On the training stage, the speech signals (training database) are pre-processed and the features are extracted. Using these features, speakers' models are estimated. A background model (Universal Background Model [Reynolds, 1995] or Cohort Models [Tran and Wagner, 2001]) is also estimated. On the testing stage of an identification system, the input, namely the speaker's utterance undergoes pre-processing and feature extraction. Then, a pattern-matching scheme finds the best-matched trained template (model) to the current unknown-speaker's utterance. On the testing stage of a verification system, there is one more input to the system, which is an identity claim. The pattern matching is made between the claimed identity trained template (model) and the current unknown speaker utterance. A true/false decision is made using a threshold measure.

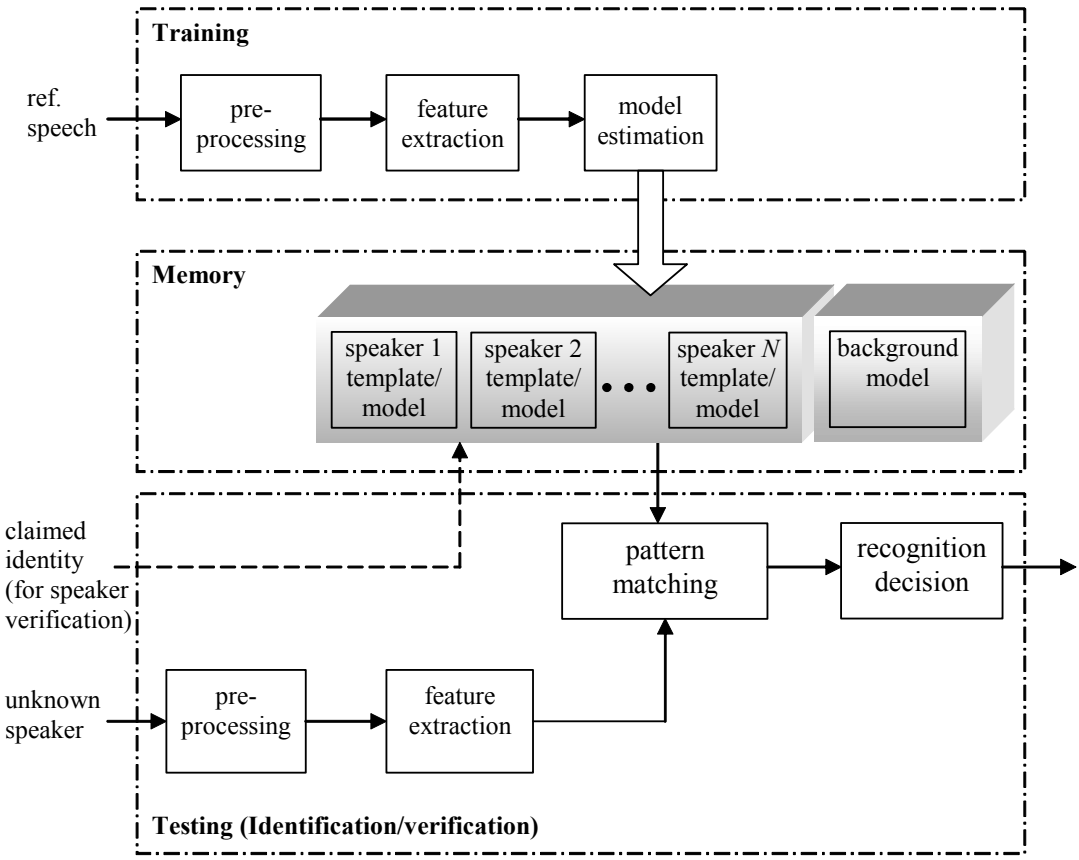

Figure 1.1: Basic speaker recognition system.

Today, Automatic Speaker Verification (ASV) systems use a common set of features (feature space) for all speakers. Most often, this common set of features consists of cepstral and delta-cepstral coefficients, which are used for speech recognition tasks. Only a few researches have been made on selecting the feature space for speaker verification/identification tasks [Chung & Eisenstein, 1978], [Cohen & Froind, 1989], [Campbell, 1992], [Charlet & Jouvet, 1997], [Pandit & Kittler, 1998], [Hayder et al., 1998], [Cohen & Zigel, 2002], [Zigel & Cohen, 2004].

The motivation for our research is the assumption that every speaker has his own 'optimal' feature space, which optimally discriminates him from other speakers. This was

supported by preliminary past work [Cohen & Froind, 1989]. The goal of this research is to demonstrate the significance of employing individual feature spaces in modern Continuous Density Hidden Markov Model (CD-HMM) [Rabiner & Juang, 1993] or Gaussian Mixture Model (GMM) [Reynolds, 1995] based verification systems, and hence, to improve verification performance. To do this, we developed criterion for feature selection, which is suitable for speaker verification tasks and correlated with the recognition rate, named "Recognition Related Criterion" (RRC). Two verification systems, which combine the individual feature selection algorithm, have been developed and implemented; the first is a text-dependent speaker verification system, based on a CD-HMM classifier and the second is a text-independent verification system, based on a GMM classifier.

Figure 1.2 shows a simple block diagram of the proposed speaker verification system. A detailed explanation is described later on chapter 4. In the training stage, the proposed speaker verification system consists of the speakers' (targets) training utterances, pre-processing, and global feature extraction; these features are stored in a general library. Algorithm for individual feature selection is executed on the feature library to yield the optimal individual feature space. For each speaker, the index of his selected features is stored and his model is trained in his own feature space, as well as his individual background model. In the testing stage, the inputs are the tested speakers' utterances and their identity claim. From the identity claim, the appropriate features list is drawn and feature extraction is made on the pre-processed utterance to yield features, which belong to the speaker feature space. The verification algorithm provides a probabilistic score which is compared to a threshold, to yield an accept or reject decision. The text-dependent system was evaluated on local high-quality text-dependent database as well as on noisy database, using several feature

selection procedures. The text-independent system was evaluated using NIST98 and NIST99 database [Martin and Przybocki, 2000].

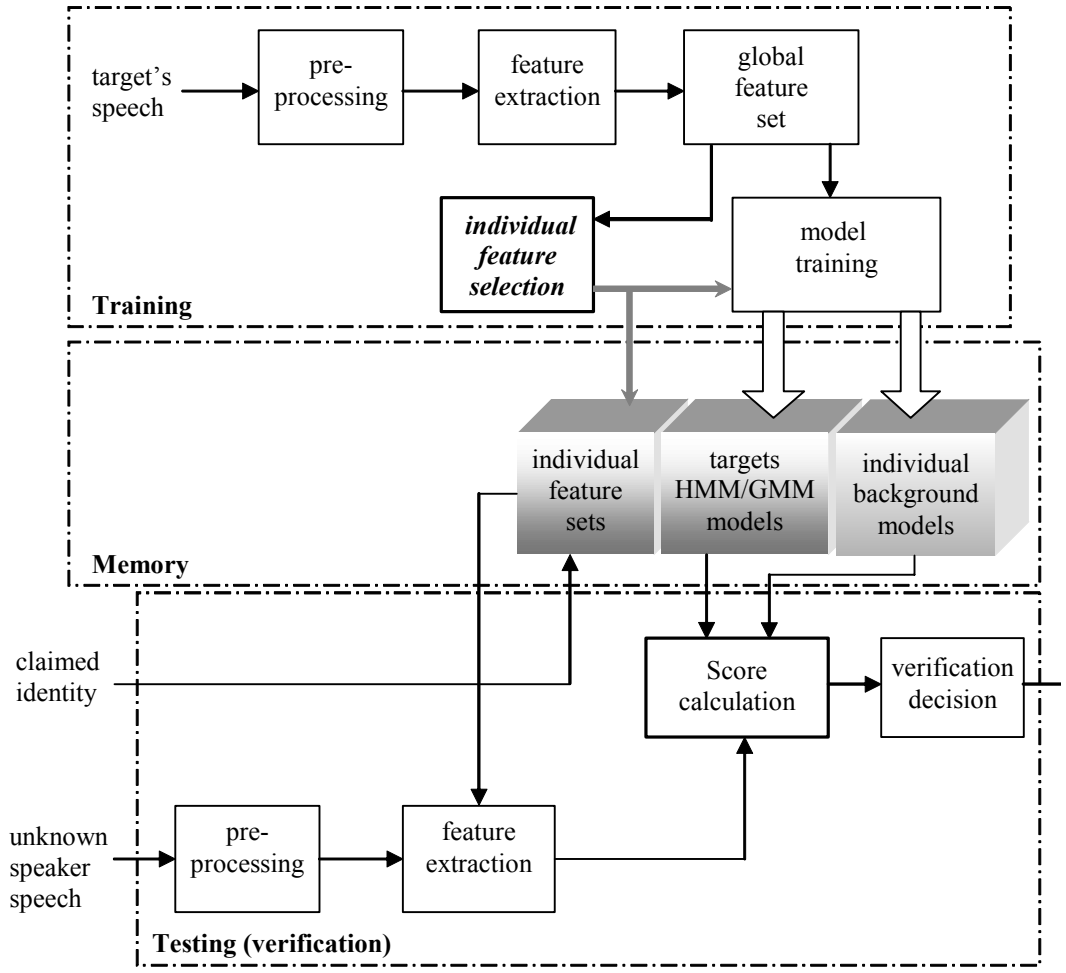

Figure 1.2: A basic block diagram of the proposed speaker verification system using individual feature space.

This thesis contains six chapters: **Chapter 1** - introduction. **Chapter 2** reviews the speaker recognition background. **Chapter 3** reviews feature selection in general and presents the proposed feature selection criteria for speaker recognition. **Chapter 4** describes the proposed speaker verification system**. Chapter 5** presents the experiments and their results. **Chapter 6** presents the conclusions and recommendations for future work.

2. Speaker Recognition Background

2.1. The Speech Wave

The speech waveform is an acoustic sound pressure wave that originates from voluntary movements of anatomical structures that comprise the human speech production system. Figure 2.1 portrays a *medium saggital* section of the speech system.

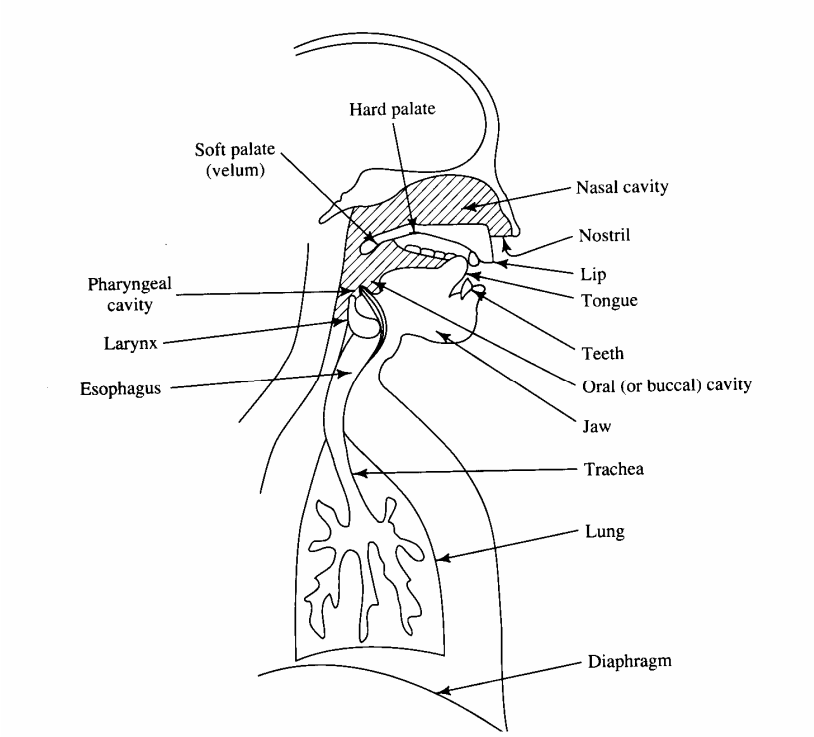

Figure. 2.1: A schematic diagram of the human speech production mechanism [Deller et al., 1993]

The gross components of the system are the *lungs*, *trachea* (windpipe), *larynx* (voice folds), *pharyngeal cavity* (throat), *oral* or *buccal cavity* (mouth), and *nasal cavity* (nose). In technical discussions, the pharyngeal and oral cavities are usually grouped into one unit referred to as the *vocal tract*, and the nasal cavity is often called the *nasal tract*. The vocal tract begins at the output of the larynx, and terminates at the input to the lips. The nasal tract

begins at the velum (which controls its opening and closure) and ends at the nostrils of the nose. Finer anatomical features critical to speech production include the *vocal folds* or *vocal cords*, *soft palate* or *velum*, *tongue*, *teeth*, and *lips*. These components move to different positions to produce various speech sounds and are known as *articulators*.

It is useful for engineers to think of speech production in terms of an acoustic filtering operation (technical model). The three main cavities of the speech production system (oral, pharyngeal, and nasal tracts) comprise the main acoustic filter. The filter is excited by the organs below it and is loaded at its main output by a radiation impedance due to the lips. The articulators, most of which are associated with the filter itself, are used to change the properties of the system, its form of excitation, and its output loading over time. A simplified acoustic model is shown in Figure 2.2.

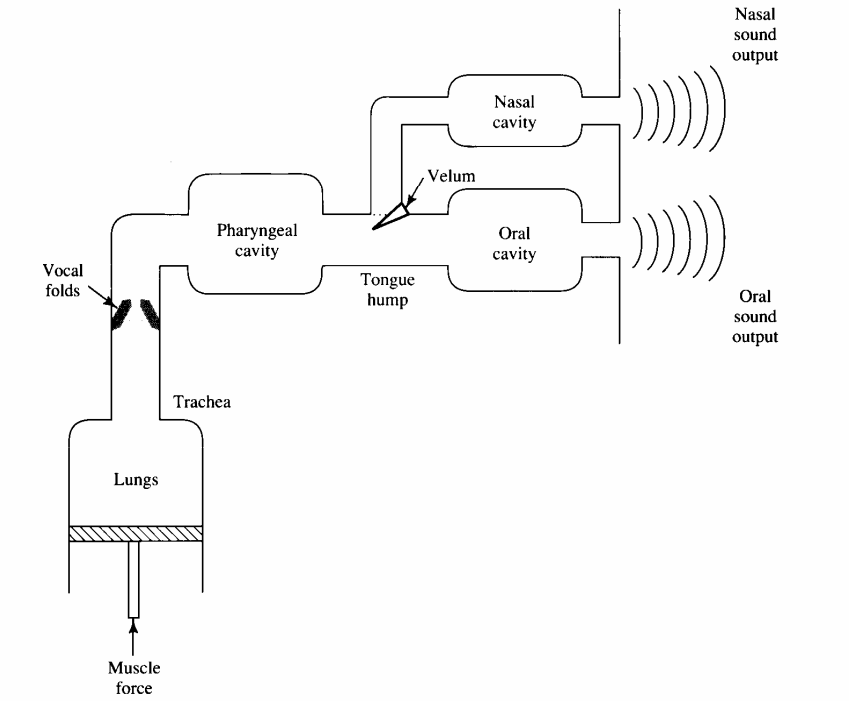

Figure 2.2: A block diagram of human speech production [Deller et al., 1993].

The main cavities (acoustic filter) contribute to the resonant structure of human speech. Repositioning of the vocal tract articulators causes the cross-sectional area of the vocal tract to vary along its length from zero (complete closure) to greater than $20cm^2$.

From a technical point of view, the larynx has a simple, but highly significant, role in speech production. Its function is to provide a periodic excitation to the system for speech sounds, which are called "voiced". The periodic vibration of the vocal folds is responsible for this voicing.

The spectral characteristics of the speech wave are time-varying (nonstationary), since the physical system changes rapidly over time. As a result, speech can be divided into sound segments that possess similar acoustic properties over short periods of time. Speech sounds are typically partitioned into two broad categories: (1) vowels that contain no major airflow restriction through the vocal tract, and (2) consonants that involve a significant restriction and are therefore weaker in amplitude and often "noisier" than vowels. Figure 2.3 shows an example of a speech wave for the word "six" being spoken by a male speaker.

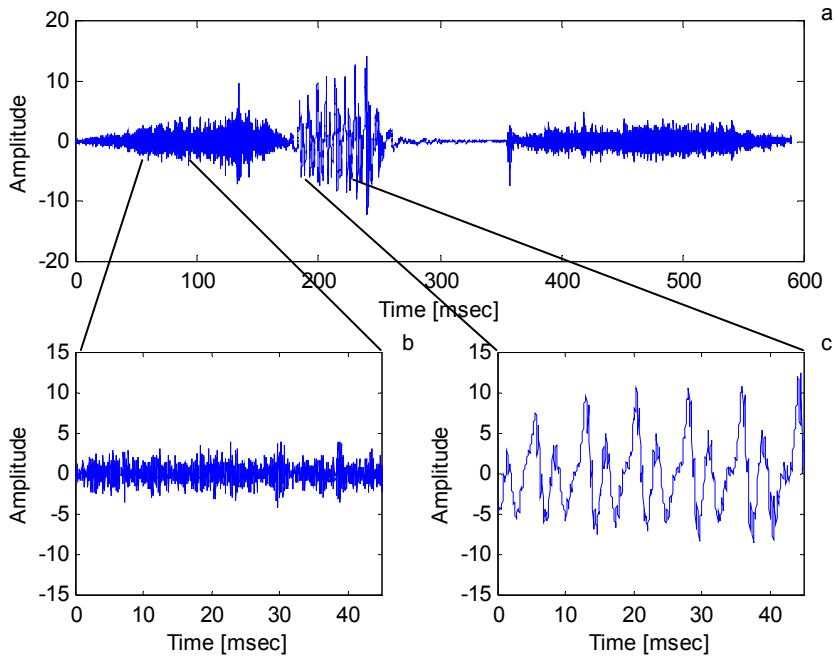

Figure 2.3: (a) A speech signal of the word "six"; (b) blowup of the initial /s/; (c) blowup of the vowel /i/.

The leading consonant sound "s" (/s/) is low in amplitude and noiselike. The subsequent vowel portion (/i/) is higher in amplitude and contains a strong periodic structure. The interplay between sounds in an utterance is called *coarticulation* [Deller et al., 1993].

Due to the limitation of the organs for human speech production and the auditory system, typical human speech communication is limited to a bandwidth of 7-8 kHz. For both sounds (especially for the vowel sound), there are well-defined regions of emphasis ("resonances") and deemphasis ("antiresonances") in the spectrum. The locations of these resonances in the frequency domain depend upon the shape and physical dimensions of the vocal tract. These resonances are called *formants*. In principle, there are an infinite number of formants in a given sound, but in practice, we usually find 3-5 (F_1, F_2, F_3, …) in the Nyquist band after sampling.

One of the principal features of any speech sound is the manner of excitation. The two main excitation types are: (1) voiced, and (2) unvoiced. Four other types of excitation, which are really just combination of voiced, unvoiced, and silence, are usually delineated for modeling and classification purposes. These are: (3) mixed, (4) plosive, (5) whisper, and (6) silence. One or more of these excitation types may be blended in the excitation of a particular speech sound or class of sounds. These excitation types pertain to English, Hebrew, and many other modern and ancient languages.

Voiced (phonation) sounds are produced by forcing air through the glottis or an opening between the vocal folds. The tension of the vocal cords is adjusted so that they vibrate in oscillatory fashion. The periodic interruption of the subglottal airflow results in quasi-periodic puffs of air that excite the vocal tract (for example: the vowel /I/). In this case, the excitation of the voiced system is the glottal pulse. *Unvoiced* sounds are generated by forming a constriction at some point along the vocal tract, and forcing air through the constriction to produce turbulence (for example: the /s/ sound).

There are some terms associated with the voiced sounds, such as the *fundamental frequency*, which is the rate of vibration of the vocal folds. The term *pitch* is often used interchangeably with fundamental frequency [Deller et al., 1993]. For men, the possible pitch range is usually found somewhere between 50-250Hz, while for women the range usually falls somewhere between 120-500Hz. Everyone has a “habitual pitch level”, which is a sort of preferred pitch that is used naturally on the average. Pitch is shifted up and down in speaking in response to factors relating to stress, intonation, and emotion. *Intonation* is associated with the pitch contour over time and performs several functions in language, the most important being to signal grammatical structure.

2.2. Speech Signal Pre-processing and Feature Extraction (Front-End Processing)

2.2.1. Pre-Processing

Speech pre-processing extracts the desired information from a speech signal. Figure 2.4 shows a block diagram of speech signal pre-processing and feature extraction.

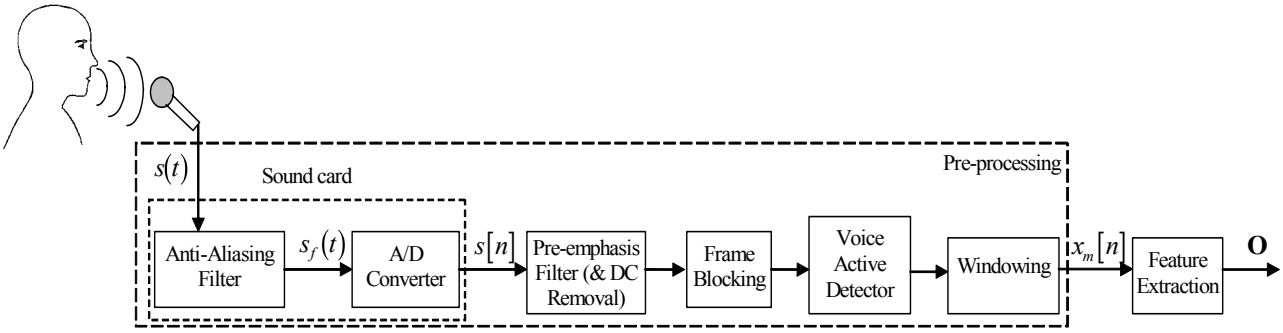

Figure 2.4: Block diagram of speech signal pre-processing and feature extraction.

Initially, the acoustic sound pressure wave is transformed into a digital signal suitable for voice processing. A microphone can be used to convert the acoustic wave into an analog electrical signal, $s(t)$. This analog signal is conditioned with *anti-aliasing filtering*. The anti-aliasing filter limits the bandwidth of the signal to approximately the Nyquist rate prior to sampling. The conditioned analog signal, $s_f(t)$, is then sampled to form a digital signal, $s[n]$, by an *analog-to-digital* converter. Today's A/D converters for speech applications typically sample with 12-16 bits of resolution at 8000-20000 samples per second.

The digitized speech signal, $s[n]$, is put through a *pre-emphasis* filter (low-order digital filter; typically a first-order FIR filter) to spectrally flatten the signal and to make it less

susceptible to finite precision effects occurring later in the signal processing. The most widely used pre-emphasis network is the fixed first-order system:

$$H(z) = 1 - \tilde{a} z^{-1} \quad ; \quad 0.9 \le \tilde{a} \le 1.0 \tag{2.1}$$

The most common value for $\tilde{a}$ is around 0.95 [Rabiner et al., 1993].

The speech signal is non-stationary. However, it may be considered "almost"-stationary in short time frames (10 – 30 msec). Therefore, the next step in processing is *frame blocking*. The pre-emphasized speech signal is blocked into frames of *N* samples, with adjacent frames being separated by *M* samples (overlapping frames; usually 50% overlap). The next step is to implement *Voice Activity Detector* (VAD) to remove non-speech frames (silence, noise,…). Usually, in high SNR applications, the VAD is based on an energy detector.

The speech frames then have to undergo *window*ing, so as to minimize the signal discontinuities at the beginning and end of each frame. Most often the Hamming window is used, which has the form

$$w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \; 0 \le n \le N-1 \tag{2.2}$$

The goal of the *feature extraction* is to effectively represent the speech frame, $x_m[n]$. This is done by extracting a single feature vector, $\mathbf{o}_m$, from the *m*th windowed frame. The sequence of feature vectors represents the speech signal, and it is the input of the speaker recognition system.

2.2.2. Feature Extraction

Speaker identity is correlated with the physiological and behavioral characteristics of the speech production system for each speaker. In order to estimate these characteristics; suitable features have to be extracted from the speech signal. Most of the features currently used for speaker recognition are associated with the spectral information of the speech frame. These features may be sorted into the following categories:

- *Static vs. Dynamic Features*: Static features are derived from a single speech frame, thus they contain no inter-frame information. Dynamic features are derived from more than one speech frame, thus they contain information also on the dynamics of the signal.

- *Direct vs. Indirect Features*: Direct features are extracted from the sampled signal without the signal being parametrically modeled. Indirect features are extracted from a spectral estimation that models the signal. The most common indirect features are based on LPC (Linear Prediction Coefficients). The LPC model the signal using Auto Regressive (AR) parametric spectral estimation. LPC analysis can be used to model the human vocal tract, and is widely used in speech coding, analysis and recognition [Mammone et al., 1996].

The following is a brief description of the features commonly used in speech/speaker recognition research. It is not intended to be an exhaustive list of all possible speech features.

2.2.2.1 Linear Prediction Coefficients

A linear model of speech production was developed by Fant in the late 1950s, in which the glottal pulse, vocal tract, and radiation are individually modeled as linear filters [Mammone et al., 1996]. A complete model of speech production represented in the z-transform domain is shown in Figure 2.5.

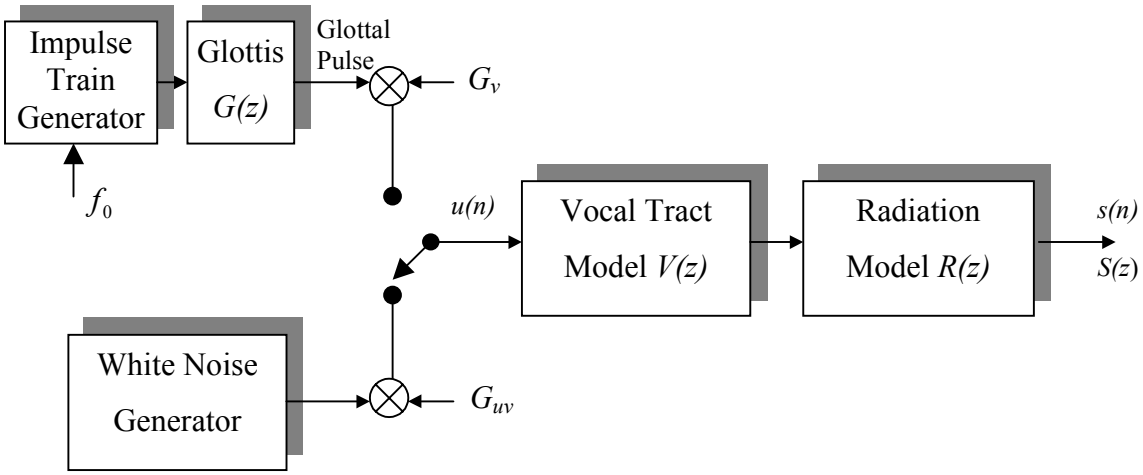

Figure 2.5: The Linear Vocal Tract Model for speech production [Deller et al., 1993].

The source is either a quasi-periodic glottal pulse sequence for the voiced sounds or a random white noise sequence for unvoiced sounds with a gain factor G_v and G_{uv}, respectively, set to control the intensity of the excitation. The transfer function *V(z)* for the vocal tract relates volume velocity at the source to pressure at the lips. It is generally an all-pole model for most speech sounds. The glottal pulse model *G(z)* describes the glottis filtering of the impulse train excitation in the voiced sounds. The radiation model *R(z)* describes the acoustical loading of the environment. The combination of the glottal pulse model and the radiation model (*G(z)R(z)*) can be reasonably approximated by a first-order backward difference. By using pre-emphasis filter in the speech pre-processing stage (equation (2.1)), the influence of these models is removed and the combination of the pre-emphasis, glottal pulse, vocal tract, and radiation yields a single all-pole transfer function given by:

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{2.3}$$

where G represents an overall gain term. With this transfer function, we get a difference equation for synthesizing the speech samples *s(n)* as

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n) \tag{2.4}$$

It can be noted that *s(n)* is predicted as a linear combination of the previous p samples. Therefore, the speech production model is often called the *linear prediction* (LP) model, or the *autoregressive model*. The coefficients $\{a_k\}$ are the *linear predictor coefficients* (LPC) and are assumed constant over the speech analysis frame.

A computationally efficient algorithm known as the Levinson-Durbin recursion [Deller et al., 1993] can be used to solve this system of equations ($r(i)$ are the autocorrelation coefficients):

$$\begin{aligned}
&E^{(0)} = r(0) \\
&a_1^{(1)} = k_1 = \frac{r(1)}{E^{(0)}} \\
&E^{(1)} = (1-k_i^2)E^{(0)} \\
&\text{for } i = 2,3,\ldots,p \\
&\quad k_i = \frac{\left\{ r(i) - \sum_{j=1}^{L-1} a_j^{(i-1)} r(|i-j|) \right\}}{E^{(i-1)}} \\
&\quad a_i^{(i)} = k_i \\
&\quad a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)} \qquad 1 \le j \le i-1 \\
&\quad E^{(i)} = (1-k_i^2)E^{(i-1)}
\end{aligned} \tag{2.5}$$

The LPC coefficients are:

$$a_i = a_i^{(p)}, \qquad 1 \le i \le p \tag{2.6}$$

2.2.2.2 PARCOR

PARCOR (partial correlation) features are also called reflection coefficients.

The vocal tract can be modeled as an electrical transmission line, a waveguide, or an analogous series of cylindrical acoustic tubes [Deller et al., 1993]. At each boundary, a portion of the wave is transmitted and the remainder is reflected. The reflection coefficients k_i are the percentage of the reflection at these discontinuities.

When Levinson-Durbin's algorithm (2.5) is used to solve the LP equations, the reflection coefficients are the intermediate k_i variables in the recursion. The reflection coefficients can also be obtained from the LP coefficients using the backward recursion [Campbell, 1997].

2.2.2.3 Log Area Ratios

Narrow bandwidth poles in the acoustic modeling of the vocal tract result in $|k_i| \approx 1$. An inaccurate representation of these PARCORs can cause gross spectral distortion. Taking the log of the area ratios results in more uniform spectral sensitivity. The *log area ratios* (LAR) are defined as

$$g_i = \log\left[\frac{A_{i+1}}{A_i}\right] = \log\left[\frac{1+k_i}{1-k_i}\right] = 2\tanh^{-1} k_i, \qquad i = 1,2,\ldots,p \tag{2.7}$$

where k_i is the *i*th PARCOR coefficient and A_i is the *i*th tube's cross-sectional area in the acoustic tube model of speech production [Campbell, 1997].

2.2.2.4 Line Spectrum Pairs

Line Spectrum Pairs (LSP) are a representation of the LPC's of the inverse filter $A(z)$, where the p zeros of $A(z)$ are mapped onto the unit circle in the z-plane through a pair of auxiliary $(p+1)$-order polynomials: $P(z)$ (symmetric) and $Q(z)$ (antisymmetric):

$$A(z) = \frac{P(z) + Q(z)}{2}$$

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1})$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) \qquad (2.8)$$

where the LSP's are the frequencies ($\omega_i ; i = 1 \ldots p$) of the zeros of $P(z)$ and $Q(z)$ [Deller et al., 1993]. The LSP satisfy an interlacing property of the zeros of the P and Q polynomials, which holds for all minimum phase $A(z)$ polynomials:

$$0 = \omega_0^{(Q)} < \omega_1^{(p)} < \omega_2^{(Q)} < \ldots < \omega_{p-1}^{(p)} < \omega_p^{(Q)} < \omega_{p+1}^{(P)} = \pi \qquad (2.9)$$

Each complex zero of $A(z)$ maps into one zero in each $P(z)$ and $Q(z)$. When the $P(z)$ and $Q(z)$ frequencies are close, it is likely that the original $A(z)$ zero was close to the unit circle, and a formant is likely to be in between the corresponding LSP. Distant P and Q zeros are likely to correspond to wide bandwidth zeros of $A(z)$ and most likely contribute only to shaping or spectral tilt [Campbell, 1997].

2.2.2.5 Cepstral Features

Cepstrum

The *real cepstrum* (RC) of a speech sequence $s(n)$ is defined as

$$c(n) = f^{-1}\left\{\log\left|f\left\{s(n)\right\}\right|\right\} \tag{2.10}$$

in which $f\{\cdot\}$ denotes the DTFT. Ordinarily, the natural or base 10 logarithm is used in this computation, but in principle any base can be used [Deller et al., 1993]. The *complex cepstrum* (CC) is less popular and its definition is: $c(n) = f^{-1}\left\{\log f\left\{s(n)\right\}\right\}$ where the log denotes the complex logarithm.

The LP Cepstrum

An important LPC parameter set, which can be derived directly from the LPC coefficient set, is the LPC cepstral coefficients, $c(m)$. The recursion used is:

$$\begin{aligned} c_0 &= \ln G \\ c_m &= a_m + \sum_{k=1}^{m-1}\left(\frac{k-m}{m}\right)a_k c_{m-k}, \qquad 1 \le m \le p \\ c_m &= \sum_{k=1}^{p}\left(\frac{k-m}{m}\right)a_k c_{m-k}, \qquad m > p \end{aligned} \tag{2.11}$$

where G is the estimated model gain [Deller et al., 1993].

The Mel-Cepstrum

A *mel* is a unit of measure of *perceived pitch* or *frequency* of a tone. It does not correspond linearly to the physical frequency of the tone, as the human auditory system apparently does not perceive pitch linearly. The mel scale is shown in Figure 2.6. The mapping is

approximately linear below 1 kHz and logarithmic above. Such an approximation is usually used in speech recognition [Deller et al., 1993].

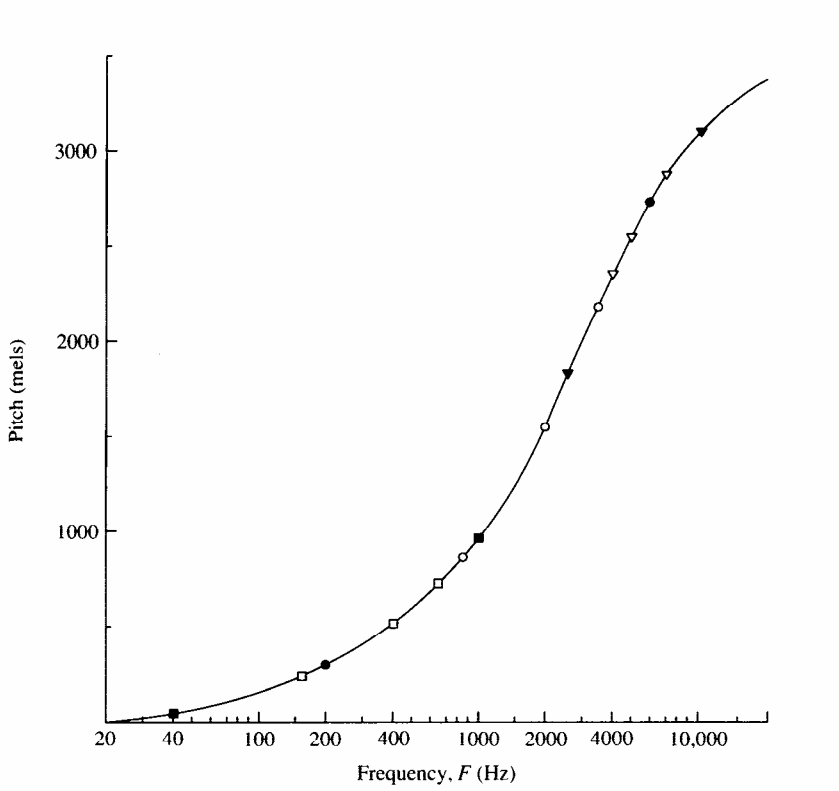

Figure 2.6: The mel scale [Deller et al., 1993].

It also has been found that the perception of a particular frequency by the auditory system is influenced by energy in a *critical band* of frequencies around it. Furthermore, the bandwidth of a critical band varies with frequency, beginning at about 100 Hz for frequencies below 1 kHz, and then increasing logarithmically above 1 kHz. Therefore, rather than simply using the mel-distributed log magnitude frequency components, some investigators have suggested using the *log total energy* in critical bands around the mel frequencies as inputs to the final IDFT.

Figure 2.7 shows an example of a filter bank (for these critical bands) in which each filter has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval. The spacing is approximately 150 mels and the width of the triangle is 300 mels [Rabiner et al., 1993].

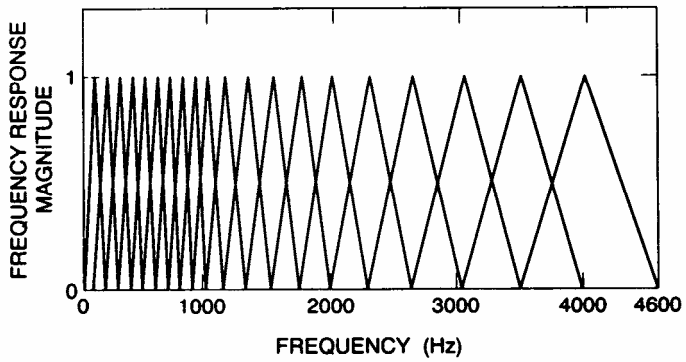

Figure 2.7: Mel Frequency Filter Bank [Rabiner et al., 1993]

The *mel-frequency cepstrum coefficients* (MFCC) are computed as:

$$MFCC_i = \sum_{k=1}^{K} X_k \cos\left[i\left(k - \frac{1}{2} \right) \frac{\pi}{K} \right], \quad i = 1,2,\ldots, M \tag{2.12}$$

where M is the number of cepstrum coefficients and $X_k, k = 1,2,\ldots,K$, represents the log-energy output of the kth filter (from K filters) [Davis et al., 1980]. Some investigators use the mel-warped cepstrum, without a filter bank [Campbell, 1997].

2.2.2.6 Dynamic Features

Dynamic features represent a discrete time derivative of the static features. These dynamic features are usually defined as:

$$\Delta c_m(n) = \frac{\sum_{i=-T}^{T} k_i c_m(n+i)}{f(k)} \qquad (2.13)$$

where $c_m[n]$ denotes the *m*th feature for the *n*th time frame, k_i is the *i*th weight and $f(k)$ is some normalization function of the weights. Delta-Delta features are also frequently used. These can be computed by applying equation (2.13) to $\Delta c_m(n)$.

2.2.2.7 Prosodic Features

Prosodic features represent the tonal and rhythmic aspects of speech. The acoustic patterns of prosodic features are heard according to systematic changes in duration, intensity, fundamental frequency, and spectral patterns of the individual phonemes [Deller et al., 1993].

From all the prosodic features, *Intonation* and *stress* convey the most linguistic information. Stress is used to distinguish similar phonetic sequences or to highlight a syllable or word against a background of unstressed syllables. Stress is measured by short time energy, and the *energy contour* can be helpful in identifying a speaker. Intonation refers to the distinctive use of patterns of pitch or melody. Intonation performs several useful functions in language, the most important being to signal grammatical structure. An analysis of intonation can be performed by considering *pitch contours*. The pitch is measured only in the voiced segments and can be estimated, for example, by the SIFT algorithm [Markel, 1972]. Pitch statistics and pitch contours in utterances can be used to characterize a person. Both short time energy and pitch exhibit large intra-speaker variability and are influenced by the subject's mood; thus they cannot be used by themselves as reliable features [Campbell, 1992].

2.3. Principles and Methods of Speaker Recognition

In chapter 1, there is a basic description of speaker recognition systems, verification vs. identification, and text-dependent vs. text-independent methods. There are some important constraints, which have to be taken into consideration for designing a speaker recognition system:

Natural constraints - the human voice is not static as are other intrinsic characteristics and it is highly sensitive to the speaker's physiologic and psychological states. Environment conditions (heat, coldness), psychological stress, hoarseness, etc. change speech characteristics, making the recognition task more difficult. Moreover, the voice varies slowly with age; therefore, there is a need for model retraining (manually or adaptively) every several months.

Environment constraints - for most of applications, the environment varies for every system login: different SNRs, channel transmission (telephone, internet), and microphone transmission (different microphones/handsets, distances, angles).

Application constraints - the number of speaker training repetitions is very limited because the user can not be too bothered. Moreover, there is no capability to train a speaker every time period (three months).

In verification systems, as well as in open-set identification, there are two types of errors to be considered for performance evaluation: the *False Acceptance* (FA) of an invalid user and the *False Rejection* (FR) of a valid user (sometimes referred as miss probability). FA and FR errors are not independent. It is usually possible to reduce one by increasing the other. Often, the Receiver Operating Characteristics (ROC) curve or Detection Error Trade-off (DET) curve are used to show the relations between these two errors in a given system.

In speaker verification, the utterance (features) of the unknown speaker is compared with the model of the speaker whose identity is claimed. If the match is good enough (above a certain threshold), the identity claim is accepted. A high threshold makes it difficult for impostors to be accepted by the system (low FA), but at the risk of falsely rejecting valid users (high FR). Conversely, a low threshold enables valid users to be accepted consistently (low FR), but at the risk of accepting impostors (high FA). To set the threshold at the desired level, it is necessary to know the distribution of target and impostor scores. The Equal-Error Rate (EER) is a commonly accepted scalar overall measure of speaker verification system performance. It corresponds to the threshold at which the false acceptance rate is equal to the false rejection rate.

An important factor to be taken into consideration when designing a text-independent recognition system is the utterance duration limitation. If there is no severe limitation on the utterance, average features can be taken from at least 15-second segments. If there is severe limitation on the utterance duration, one has to build phoneme or PLU (Phoneme Like Units) models for each speaker, and speaker recognition system based speech recognition is made (see next section).

2.3.1. Speaker Recognition Methods

Many pattern-matching methods for speaker recognition (identification and verification) have been proposed in the literature over the last 30 years. These methods involve computing a match score, which is a measure of the similarity of the input feature vectors to some model. Speaker models are constructed from the features extracted from the speech signal. To enroll users into the system, a model of the voice, based on the extracted features, is

generated and stored. To recognize a user, the matching algorithm compares/scores the incoming speech signal with the model of the claimed user.

There are two types of models: *template models* and *stochastic models*. In stochastic models (e.g. HMM, GMM), the pattern matching is probabilistic and results in a measure of the likelihood, or conditional probability, of the observation given the model. For template models (e.g. DTW, VQ), the pattern matching is deterministic. The observation is assumed to be an imperfect replica of the template, and the alignment of observed frames to template frames is selected to minimize a distance measure *d* [Campbell, 1997].

The design of speaker recognition systems is highly dependent on the application. Thus, systems for speaker recognition are usually divided to text-dependent and text-independent speaker recognition methods [Furui, 1997].

2.3.1.1 DTW Based Methods

A typical approach to text-dependent speaker recognition is to perform spectral template matching with Dynamic Time Warping (DTW). In this approach, each utterance is represented by a sequence of feature vectors, generally, short-term spectral feature vectors. Trial-to-trial timing variation of utterances of the same text is normalized by aligning the analyzed feature vector sequence of a test utterance to the template feature vector sequence using a DTW algorithm. The overall distance between the test utterance and the template (reference) is used for the recognition decision. An example of DTW system for text-dependent speaker recognition can be found in [Furui, 1981].

2.3.1.2 Long-Term Statistics-Based Methods

Long-term sample statistics (such as mean and variance) of various spectral features over a series of utterances have been used for text-independent tasks [Markel et al., 1977]. Long-term statistics-based methods are effective for long utterances. Utterances with a duration of over 15-second long, have a phoneme distribution quite similar to natural language distribution. An example for such method exists in [Markel et al., 1979]. [Campbell, 1997] used the mean and covariance of feature vectors, from voiced segments alone. He investigated several distance measures and features.

2.3.1.3 VQ-Based Methods

VQ-based methods are more suitable for coping with short-term utterances than long-term statistics-based methods. This is an unsupervised clustering method by which VQ codebooks consisting of a small number of representative feature vectors can efficiently characterize speaker specific features [Li et al., 1983] [Rosenberg et al., 1987]. A speaker-specific codebook is generated by clustering the training feature vectors of each speaker. In the recognition stage, an input utterance is vector-quantized by using the codebook of each reference speaker and then the VQ distortion accumulated over the entire input utterance is used for making the recognition determination (see Figure 2.8 below).

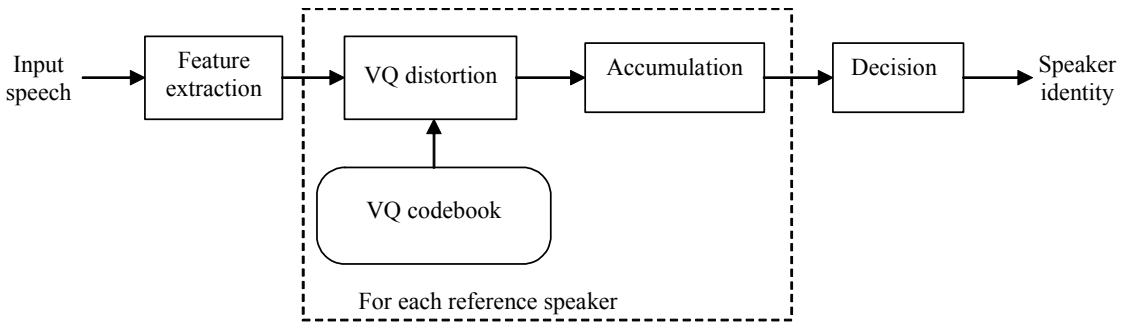

Figure 2.8: VQ based method.

2.3.1.4 HMM Based Methods

The Hidden Markov Model (HMM) is a stochastic function of a Markov chain (see figure 2.9). As such it is composed of two elements: a Markov process and a set of stochastic functions, or output probabilities [Rabiner et al., 1993].

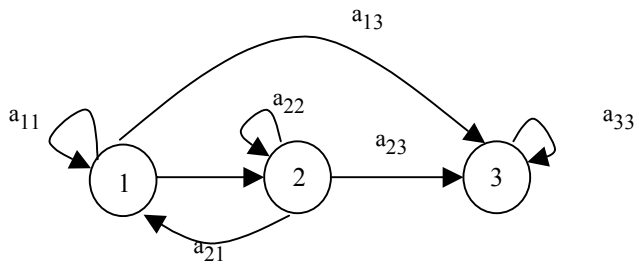

Figure 2.9: HMM structure.

There are two kinds of output probabilities to consider: discrete probability which is used in Discrete Density HMM (DDHMM), and Continuous probability which is used in Continuous Density HMM (CDHMM). Two kinds of model architectures are often used in speaker recognition systems: *left-to-right model*, where the transitions between states are constrained to the left-to-right direction (suitable for text-dependent tasks). An HMM can efficiently model the statistical variation of spectral features. Therefore, HMM-based methods have achieved better recognition accuracies than DTW-based methods [Naik et al., 1989] [Rosenberg et al., 1991] [Zheng et al., 1988].

The second architecture is *Ergodic*, for which all the state transitions are allowed (suitable for text-independent tasks). The basic system structure is the same as the VQ-based method (figure 2.8), but in this method an ergodic HMM is used instead of a VQ codebook [Poritz, 1982], [Matsui et al., 1992].

Detailed description of the HMM is described in the Appendix.

2.3.1.5 GMM-Based Methods

The Gaussian Mixture Model (GMM) is simply a one state case of the CD-HMM. A VQ-based method can be regarded as a special (degenerate) case of a GMM with a distortion measure being used as the observation probability. [Reynolds et al., 1995] proposed a text-independent speaker identification system based on GMM. State of the art text-independent speaker recognition systems use GMM based methods.

2.3.1.6 ANN-Based Methods

Artificial Neural Networks (ANN) [Haykin, 1994] learn complex mappings between inputs and outputs. ANNs include a large family of methods. Although these methods differing, all are built from simple interconnected units that cooperate to implement the global transfer function of the ANN. The functional form of the net is specified by the ANN architecture and dynamics. ANNs enable the construction of complex systems with non-linear transfer functions and sophisticated dynamics. ANN have been playing an increasing role in speaker recognition at the last 10 years [Bennani et al., 1995], [Fakotakis et al., 1999].

2.3.1.7 Speech Recognition-Based Methods

Speech recognition based methods are used to recognize phoneme-classes or phonemes. Then each phoneme (class) segment in the input speech can be compared with speaker models or templates corresponding to the particular phoneme (see figure 2.10). One can look at these methods as supervised clustering VQ, where the phonemes are determined as clusters [Savic et al., 1990], [Sukkar et al., 2000].

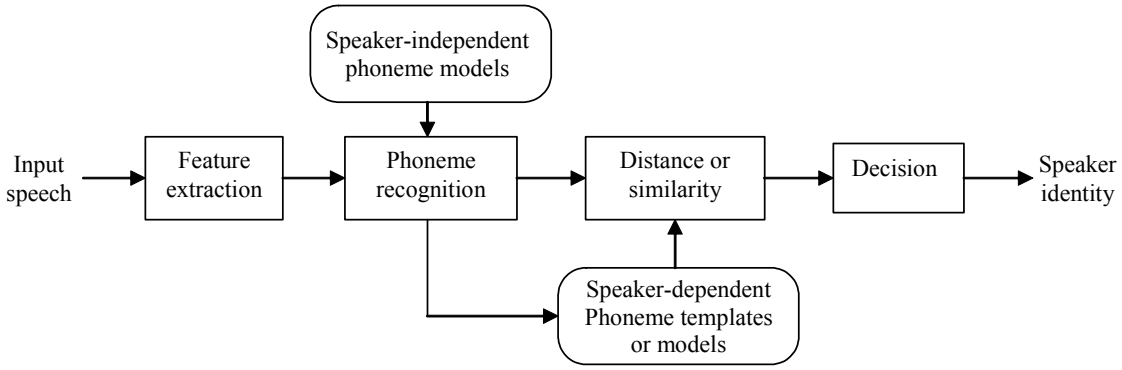

Figure 2.10: Speech-recognition-based methods.