8 SPEAKER RECOGNITION

Joseph P. Campbell, Jr.
Department of Defense
Fort Meade, MD
j.campbell@ieee.org

Abstract *A tutorial on the design and development of automatic speaker recognition systems is presented. Automatic speaker recognition is the use of a machine to recognize a person from a spoken phrase. These systems can operate in two modes: to* identify *a particular person or to* verify *a person's claimed identity. Speech processing and the basic components of automatic speaker recognition systems are shown and design tradeoffs are discussed. The performances of various systems are compared.*

Keywords: *Access control, authentication, biometrics, biomedical measurements, biomedical signal processing, biomedical transducers, communication system security, computer network security, computer security, corpus, databases, identification of persons, public safety, site security monitoring, speaker recognition, speech processing, verification.*

1. Introduction

The focus of this chapter is on facilities and network access-control applications of speaker recognition. Speech processing is a diverse field with many applications. Figure 8.1 shows a few of these areas and how speaker recognition relates to the rest of the field. This chapter will emphasize the speaker recognition applications shown in the boxes of Figure 8.1.

Speaker recognition encompasses verification and identification. Automatic speaker verification (ASV) is the use of a machine to verify a person's claimed identity from his voice. The literature abounds with different terms for speaker verification, including voice verification, speaker authentication, voice authentication, talker authentication, and talker verification. In automatic speaker identification (ASI), there is no *a priori* identity claim, and the system decides who the person is, what group the person is a member of, or (in the open-set case) that the person is unknown. General overviews of speaker recognition have been given by Atal, Doddington, Furui, O'Shaughnessy, Rosenberg, Soong, Sutherland, and Jack [2,9,13,28,38,39,46].

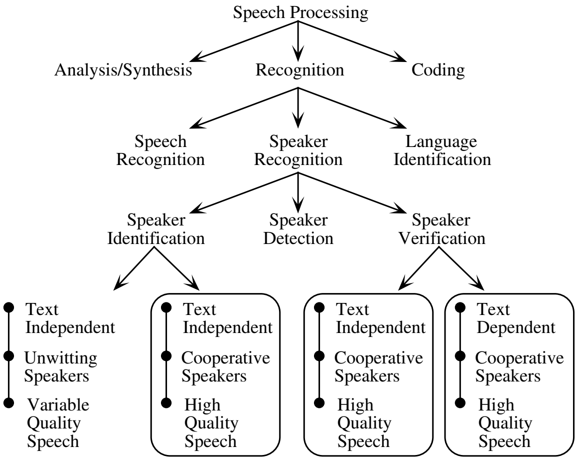

Figure 8.1 Speech processing.

Speaker verification is defined as deciding if a speaker is who he claims to be. This is different than the speaker identification problem, which is deciding if a speaker is a specific person or is among a group of persons. In speaker verification, a person makes an identity claim (e.g., entering an employee number or presenting his smart card). In text-dependent recognition, the phrase is known to the system and it can be fixed or not fixed and prompted (visually or orally). The claimant speaks the phrase into a microphone. This signal is analyzed by a verification system that makes the binary decision to accept or reject the user's identity claim or possibly to report insufficient confidence and request additional input before making the decision.

A typical ASV setup is shown in Figure 8.2. 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.

Many factors can contribute to verification and identification errors. Table 8.1 lists some of the human and environmental factors that contribute to these errors, a few of which are shown in Figure 8.2. 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.

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)

Table 8.1 Sources of verification error.

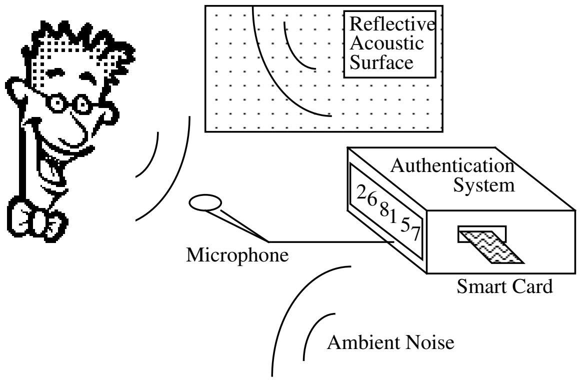

Figure 8.2 Typical speaker-verification setup.

Motivation

ASV and ASI are probably the most natural and economical methods for solving the problems of unauthorized use of computer and communications systems and multilevel access control. With the ubiquitous telephone network and microphones bundled with computers, the cost of a speaker recognition system might only be for the software for the recognition algorithm.

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 [25,36], ordinary human changes (e.g., time-of-day voice changes and minor head colds), and mimicry by humans and tape recorders [18].

Problem Formulation

Speech is a complicated signal produced as a result of several transformations occurring at several different levels: semantic, linguistic, articulatory, and acoustic. Differences in these transformations appear as differences in the acoustic properties of the speech signal. Speaker-related differences are a result of a combination of anatomical differences inherent in the vocal tract and the learned speaking habits of different individuals. In speaker recognition, all these differences can be used to discriminate between speakers.

Generic Speaker Verification

The general approach to ASV consists of five steps: digital speech data acquisition, feature extraction, pattern matching, making an accept/reject decision, and enrollment to generate speaker reference models. A block diagram of this procedure is shown in Figure 8.3. Feature extraction maps each interval of speech to a multidimensional feature space. (A speech interval typically spans 10 to 30 ms of the speech waveform and is referred to as a frame of speech.) This sequence of feature vectors $\mathbf{x}_i$ is then compared to speaker models by pattern matching. This results in a match score z_i for each vector or sequence of vectors. The match score measures the similarity of the computed input feature vectors to models of the claimed speaker or feature vector patterns for the claimed speaker. Last, a decision is made to either accept or reject the claimant according to the match score or sequence of match scores, which is a hypothesis-testing problem.

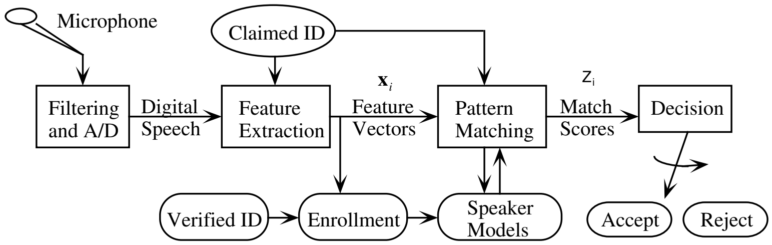

Figure 8.3 Generic speaker verification system.

For speaker recognition, features that exhibit high speaker discrimination power, high interspeaker variability, and low intraspeaker variability are desired. Many forms of pattern matching and corresponding models are possible. Pattern matching methods include dynamic time warping (DTW), hidden Markov modeling (HMM), artificial neural networks, and vector quantization (VQ). Template models are used in DTW, statistical models are used in HMM, and codebook models are used in VQ.

Previous Work

Table 8.2 shows a sampling of the chronological advancement in speaker verification. The following terms are used to define the columns in Table 8.2: “Source” refers to a citation in the references, “org” is the company or school where the work was done, “features” are the signal measurements such as linear prediction (LP) and log area ratio (LAR), “input” is the type of input speech (laboratory, office quality, or telephone), “text” indicates whether text-dependent or text-independent mode of operation is used, “method” is the heart of the pattern-matching process, “pop” is the population size of the test (number of people), and “error” is the equal error percentage for speaker verification systems “v” or the recognition error percentage for speaker identification systems “i” given the specified duration of test speech in seconds. This data is presented to give a simplified general view of past speaker-recognition research. The references should be consulted for important distinctions that are not included; e.g., differences in enrollment, differences in cross-gender impostor trials, differences in normalizing “cohort” speakers [40], differences in partitioning the impostor and cohort sets, and differences in known versus unknown impostors [5]. It should be noted that it is difficult to make meaningful comparisons between the text-dependent and the generally more difficult text-independent tasks. Text-independent approaches, such as Gish’s segmental Gaussian model [15] and Reynolds’ Gaussian Mixture Model (GMM) [36] need to deal with unique problems (e.g., sounds or articulations present in the test material, but not in training). It is also difficult to compare between the binary-choice verification task and the generally more difficult multiple-choice identification task [9,29].

There are over a dozen commercial ASV systems, including those from ITT, Lernout & Hauspie, T-NETIX, Veritel, and Voice Control Systems. Perhaps the largest scale deployment of any biometric to date is Sprint’s Voice FONCARD®, which uses TI’s voice-verification engine. Speaker verification applications include access control, telephone banking, and telephone credit cards. The accounting firm of Ernst and Young estimates that high-tech computer thieves in the U.S. steal \$3 to \$5 billion annually. Automatic speaker-recognition technology could substantially reduce this crime by reducing these fraudulent transactions. It takes a pair of subjects to make a false acceptance error: an impostor and a target. Because of this hunter and prey relationship, in this work, the impostor is referred to as a wolf and the target as a sheep. False acceptance errors are the ultimate concern of high-security speaker-verification applications; however, they can be traded off for false rejection errors.

The following section contains an overview of digital signal acquisition, speech production, speech signal processing, and speaker characterization based on linear prediction and mel cepstra modeling.

2. Speech processing

Speech processing extracts the desired information from a speech signal. To process a signal by a digital computer, the signal must be represented in digital form so that it can be used by a digital computer.

Speech Signal Acquisition

Initially, the acoustic sound pressure wave is transformed into a digital signal suitable for voice processing. A microphone or telephone handset can be used to convert the acoustic wave into an analog signal. This analog signal is conditioned with antialiasing filtering (and possibly additional filtering to compensate for any channel impairments). The antialiasing filter limits the bandwidth of the signal to approximately the Nyquist rate (half the sampling rate) before sampling. The conditioned analog signal is then sampled to form a digital signal by an analog-to-digital (A/D) converter. Today’s A/D converters for speech applications typically

Source	Org	Features	Method	Input	Text	Pop	Error
Atal [1]	AT&T	Cepstrum	Pattern Match	Lab	Dependent	10	i: 2%@0.5s v: 2%@1s
Markel and Davis [26]	STI	LP	Long Term Statistics	Lab	Independent	17	i: 2%@39s
Furui [12]	AT&T	Normalized Cepstrum	Pattern Match	Telephone	Dependent	10	v: 0.2%@3s
Schwartz, et al. [43]	BBN	LAR	Nonparametric pdf	Telephone	Independent	21	i: 2.5%@2s
Li and Wrench [23]	ITT	LP, Cepstrum	Pattern Match	Lab	Independent	11	i: 21%@3s i: 4%@10s
Doddington [9]	TI	Filter-bank	DTW	Lab	Dependent	200	v: 0.8%@6s
Soong, et al. [44]	AT&T	LP	VQ (size 64) Likelihood Ratio distortion	Telephone	10 isolated digits	100	i: 5%@1.5s i: 1.5%@3.5s
Higgins and Wohlford [19]	ITT	Cepstrum	DTW Likelihood Scoring	Lab	Independent	11	v: 10%@2.5s v: 4.5%@10s
Attili, et al. [3]	RPI	Cepstrum, LP, Autocorr	Projected Long Term Statistics	Lab	Dependent	90	v: 1%@3s
Higgins, et al. [18]	ITT	LAR, LP-Cepstrum	DTW Likelihood Scoring	Office	Dependent	186	v: 1.7%@10s
Tishby [47]	AT&T	LP	HMM (AR mix)	Telephone	10 isolated digits	100	v: 2.8%@1.5s v: 0.8%@3.5s
Reynolds [34]; Reynolds and Carlson [35]	MIT-LL	Mel-Cepstrum	HMM (GMM)	Office	Dependent	138	i: 0.8%@10s v: 0.12%@10s
Che and Lin [7]	Rutgers	Cepstrum	HMM	Office	Dependent	138	i: 0.56% @2.5s i: 0.14%@10s v: 0.62% @2.5s

Table 8.2 Selected chronology of speaker-recognition progress.

Source	Org	Features	Method	Input	Text	Pop	Error
Tishby [47]	AT&T	LP	HMM (AR mix)	Telephone	10 isolated digits	100	v: 2.8%@1.5s v: 0.8%@3.5s
Colombi, et al. [8]	AFIT	Cep, Eng dCep, ddCep	HMM monophone	Office	Dependent	138	i: 0.22%@10s v: 0.28%@10s
Reynolds [37]	MIT-LL	Mel-Cepstrum, Mel-dCepstrum	HMM (GMM)	Telephone	Independent	416	v: 11%/16% @3s v: 6%/8% @10s v: 3%/5% @30s matched/ mismatched handset

Table 8.2 Selected chronology of speaker-recognition progress (contd.).

sample with 12 to 16 bits of resolution at 8,000 to 20,000 samples per second. Oversampling is commonly used to allow a simpler analog antialiasing filter and to control the fidelity of the sampled signal precisely (e.g., sigma-delta converters).

In local speaker-verification applications, the analog channel is simply the microphone, its cable, and analog signal conditioning. Thus, the resulting digital signal can be very high quality, lacking distortions produced by transmission of analog signals over telephone lines.

YOHO Speaker-Verification Corpus

The work presented here is based on high-quality signals for benign-channel speaker verification applications. The primary database for this work is known as the YOHO Speaker Verification Corpus, which was collected by ITT under a U.S. Government contract. The YOHO database was the first large-scale, scientifically controlled and collected, high-quality speech database for speaker-verification testing at high confidence levels. Table 8.3 describes the YOHO database [17]. YOHO is available from the Linguistic Data Consortium (University of Pennsylvania) and test plans have been developed for its use [5]. This database already is in digital form, emulating the third generation Secure Terminal Unit's (STU-III) secure voice telephone input characteristics, so the first signal processing block of the verification system in Figure 8.3 (signal conditioning and acquisition) is taken care of.

In a text-dependent speaker-verification scenario, the phrases are known to the system (e.g., the claimant is prompted to say them). The syntax used in the YOHO database is "combination lock" phrases. For example, the prompt might read: "Say: twenty-six, eighty-one, fifty-seven."

YOHO was designed for U.S. Government evaluation of speaker-verification systems in "office" environments. In addition to office environments, there are enormous consumer markets that must contend with noisy speech (e.g., telephone services) and far-field microphones (e.g., computer access).

“Combination lock” phrases (e.g., “twenty-six, eighty-one, fifty-seven”)
138 subjects: 106 males, 32 females
Collected with a STU-III electret-microphone telephone handset over 3 month period in a real-world office environment
4 enrollment sessions per subject with 24 phrases per session
10 verification sessions per subject at approximately 3-day intervals with 4 phrases per session
Total of 1380 validated test sessions
8 kHz sampling with 3.8 kHz analog bandwidth (STU-III like)
1.2 gigabytes of data

Table 8.3 The YOHO corpus [5].

Speech Production

There are two main sources of speaker-specific characteristics of speech: physical and learned. Vocal tract shape is an important physical distinguishing factor of speech. The vocal tract is generally considered as the speech production organ above the vocal folds. As shown in Figure 8.4 [11], this includes the following: laryngeal pharynx (beneath epiglottis), oral pharynx (behind the tongue, between the epiglottis and velum), oral cavity (forward of the velum and bounded by the lips, tongue, and palate), nasal pharynx (above the velum, rear end of nasal cavity), and the nasal cavity (above the palate and extending from the pharynx to the nostrils). An adult male vocal tract is approximately 17 cm long [11].

The vocal folds (formerly known as vocal cords) are shown in Figure 8.4. The larynx is composed of the vocal folds, the top of the cricoid cartilage, the arytenoid cartilages, and the thyroid cartilage (also known as “Adam’s apple”). The vocal folds are stretched between the thyroid cartilage and the arytenoid cartilages. The area between the vocal folds is called the glottis.

As the acoustic wave passes through the vocal tract, its frequency content (spectrum) is altered by the resonances of the vocal tract. Vocal tract resonances are called *formants.* Thus, the vocal tract shape can be estimated from the spectral shape (e.g., formant location and spectral tilt) of the voice signal.

Voice verification systems typically use features derived only from the vocal tract. As seen in Figure 8.4, the human vocal mechanism is driven by an excitation source, which also contains speaker-dependent information. The excitation is generated by airflow from the lungs, carried by the trachea (also called the “wind pipe”) through the vocal folds (or the arytenoid cartilages). The excitation can be characterized as phonation, whispering, frication, compression, vibration, or a combination of these.

For other aspects of speech production that could be useful for speaker recognition, please refer to [6].

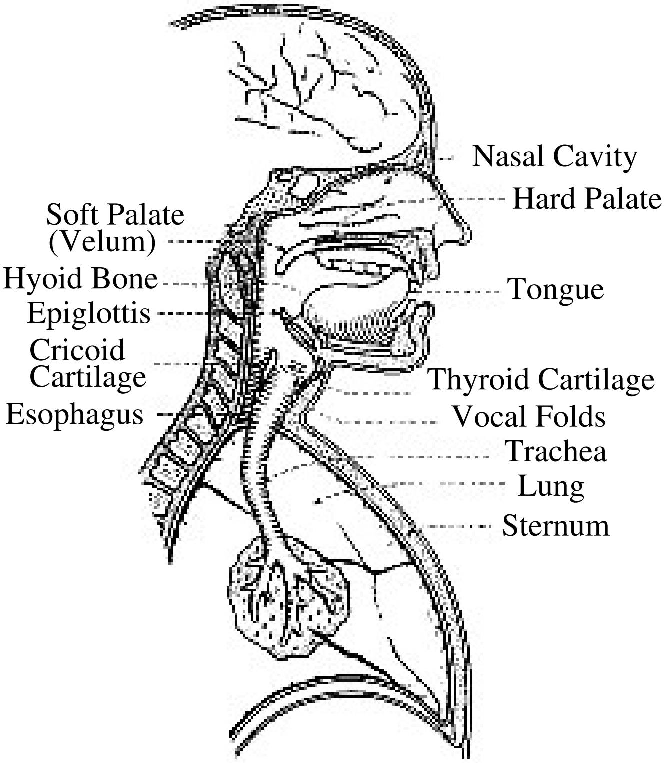

Figure 8.4 Human vocal system (reprinted with permission from J. Flanagan, *Speech Analysis Synthesis and Perception*, 2nd ed. New York and Berlin: Springer-Verlag, 1972, p. 10, Fig. 2.1 © Springer-Verlag).

Linear Prediction

The all-pole LP models a signal s_n by a linear combination of its past values and a scaled present input [24]

$$s_n = -\sum_{k=1}^{p} a_k \cdot s_{n-k} + G \cdot u_n \tag{8.1}$$

where s_n is the present output, p is the prediction order, a_k are the model parameters called the predictor coefficients (PCs), s_{n-k} are past outputs, G is a gain scaling factor, and u_n is the present input. In speech applications, the input u_n is generally unknown, so it is ignored. Therefore, the LP approximation $\hat{s}_n$, depending only on past output samples, is

$$\hat{s}_n = -\sum_{k=1}^{p} a_k \cdot s_{n-k} \tag{8.2}$$

The source u_n, which corresponds to the human vocal tract excitation, is not modeled by these PCs. It is certainly reasonable to expect that some speaker-dependent characteristics are present in this excitation signal (e.g., fundamental frequency). Therefore, if the excitation signal is ignored, valuable speaker-verification discrimination information could be lost.

Defining the prediction error e_n (also known as the residual) as the difference between the actual value s_n and the predicted value $\hat{s}_n$ yields

$$e_n = s_n - \hat{s}_n = s_n + \sum_{k=1}^{p} a_k \cdot s_{n-k} \tag{8.3}$$

Using the a_k model parameters, Eq. (8.4) represents the fundamental basis of LP representation. It implies that *any* signal is defined by a linear predictor and the corresponding LP error. Obviously, the residual contains all the information not contained in the predictor coefficients (PCs).

$$s_n = -\sum_{k=1}^{p} a_k \cdot s_{n-k} + e_n \tag{8.4}$$

From Eq. (8.1), the LP transfer function is defined as

$$H(z) \equiv \frac{S(z)}{U(z)} \equiv \frac{Z[s_n]}{Z[u_n]} \tag{8.5}$$

which yields

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} \equiv \frac{G}{A(z)} \tag{8.6}$$

where $A(z)$ is known as the p^th-order inverse filter.

LP analysis determines the PCs of the inverse filter $A(z)$ that minimize the prediction error e_n in some sense. Typically, the mean square error (MSE) is minimized because it allows a simple, closed-form solution of the PCs. For example, an 8th-order 8 kHz LP analysis of the vowel /U/ (as in "foot") had the predictor coefficients shown in Table 8.4.

Power of z	0	–1	–2	–3	–4	–5	–6	–7	–8
Predictor Coefficient	1	–2.346	1.657	–0.006	0.323	–1.482	1.155	–0.190	–0.059

Table 8.4 Example of 8th-order linear predictor coefficients for the vowel /U/ as in "foot".

Evaluating the magnitude of the *z* transform of H(*z*) at equally spaced intervals on the unit circle yields the following power spectrum having formants (vocal tract resonances or spectral peaks) at 390, 870, and 3040 Hz (Figure 8.5). These resonance frequencies are in agreement with the Peterson and Barney formant frequency data for the vowel /U/ [33].

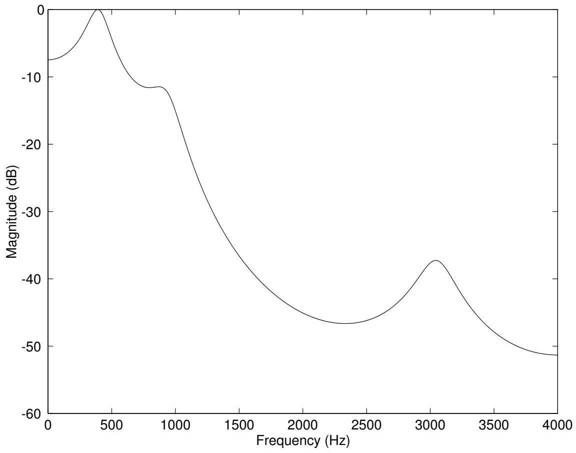

Figure 8.5 Frequency response for the vowel /U/.

Features are constructed from the speech model parameters; for example, the a_k shown in Eq. (8.6). These LP coefficients are typically nonlinearly transformed into perceptually meaningful domains suited to the application. Some feature domains useful for speech coding and recognition include reflection coefficients (RCs); log-area ratios (LARs) or arcsin of the RCs; line spectrum pair (LSP) frequencies [4,6,21,22,41]; and the LP cepstrum [33].

Reflection Coefficients and Log Area Ratios

The vocal tract can be modeled as an electrical transmission line, a waveguide, or an analogous series of cylindrical acoustic tubes. At each junction, there can be an impedance mismatch or an analogous difference in cross-sectional areas between tubes. At each boundary, a portion of the wave is transmitted and the remainder is reflected (assuming lossless tubes). The reflection coefficients k_i are the percentage of the reflection at these discontinuities. If the acoustic tubes are of equal length, the time required for sound to propagate through each tube is equal (assuming planar wave propagation). Equal propagation times allow simple *z* transformation for digital filter simulation. For example, a series of five acoustic tubes of equal lengths with cross-sectional areas A_1, …, A_5 is shown in Figure 8.6. This series of five tubes represents a fourth-order system that might fit a vocal tract minus the nasal cavity. The reflection coefficients are determined by the ratios of the adjacent cross-sectional areas with appropriate boundary conditions [33]. For a pth-order system, the boundary conditions given in Eq. (8.7) correspond to a closed glottis (zero area) and a large area following the lips.

$$
\begin{aligned}
&A_0 = 0\\
&A_{p+1} >> A_p\\
&\mathrm{k}_i = \frac{A_{i+1} - A_i}{A_{i+1} + A_i} \quad \mathit{for} \quad i = 1,2,\ldots p
\end{aligned}
\tag{8.7}
$$

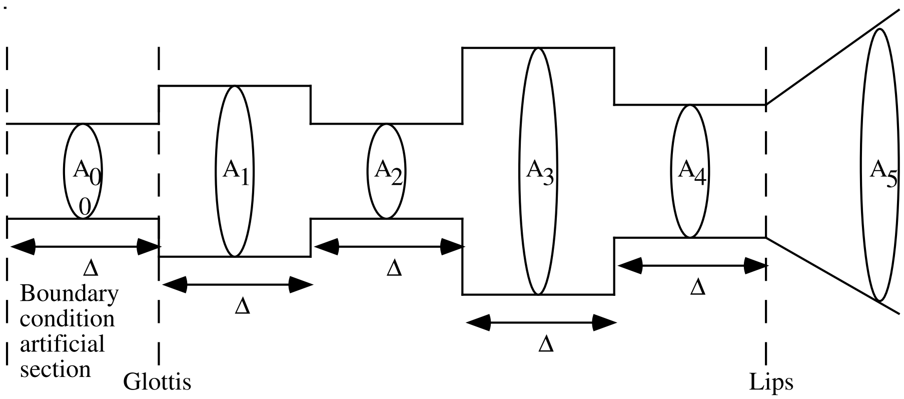

Figure 8.6 Acoustic tube model of speech production.

Narrow bandwidth poles result in $|k_i| \approx 1$. An inaccurate representation of these RCs can cause gross spectral distortion. Taking the log of the area ratios results in more uniform spectral sensitivity. The LARs are defined as the log of the ratio of adjacent cross-sectional areas

$$
g_i = \log\left[\frac{A_{i+1}}{A_i}\right] = \log\left[\frac{1+k_i}{1-k_i}\right] = 2\tanh^{-1} k_i \quad \mathit{for} \quad i = 1,2,\ldots p \tag{8.8}
$$

Mel-Warped Cepstrum

The mel-warped cepstrum is a very popular feature domain that does not require LP analysis. It can be computed as follows: 1) window the signal, 2) take the fast Fourier transform (FFT), 3) take the magnitude, 4) take the log, 5) warp the frequencies according to the mel scale, and 6) take the inverse FFT. A variation on the cepstrum is the LP-cepstrum, where steps 1 – 3 are replaced by the magnitude spectrum from LP analysis. The mel warping transforms the frequency scale to place less emphasis on high frequencies. It is based on the nonlinear human perception of the frequency of sounds [32]. The cepstrum can be considered as the spectrum of the log spectrum. Removing its mean reduces the effects of linear time-invariant filtering (e.g., channel distortion). Often, the time derivatives of the mel cepstra (also known as delta cepstra) are used as additional features to model trajectory information. The cepstrum's density has the benefit of being modeled well by a linear combination of Gaussian densities as used in the Gaussian Mixture Model [36]. Perhaps the most compelling reason for using the mel-warped cepstrum is that it has been demonstrated to work well in speaker-recognition systems [15] and, somewhat ironically, in speech-recognition systems [32], too. Furui addresses this irony and other issues plaguing speaker recognition in his set of open questions [14].

The next section presents feature selection, estimation of mean and covariance, divergence, and Bhattacharyya distance. It is highlighted by the development of the divergence shape measure and the Bhattacharyya distance shape.

3. Feature selection and measures

To apply mathematical tools without loss of generality, the speech signal can be represented by a sequence of feature vectors. The selection of appropriate features and methods to estimate (extract or measure) them are known as feature selection and feature extraction, respectively.

Traditionally, pattern-recognition paradigms are divided into three components: feature extraction and selection, pattern matching, and classification. Although this division is convenient from the perspective of designing system components, these components are not independent. The false demarcation among these components can lead to suboptimal designs because they all interact in real-world systems.

In speaker verification, the goal is to design a system that minimizes the probability of verification errors. Thus, the underlying objective is to discriminate between the given speaker and all others. A comprehensive review of discriminant analysis is given in [16]. For an overview of the feature selection and extraction methods, please refer to [6]. The next section introduces pattern matching.

4. Pattern matching

The pattern-matching task of speaker verification involves computing a match score, which is a measure of the similarity between the input feature vectors and 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 (possibly on an encrypted smart card). Then, to authenticate a user, the matching algorithm compares/scores the incoming speech signal with the model of the claimed user.

There are two types of models: stochastic models and template models. In stochastic models, 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, 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. The likelihood L can be approximated in template-based models by exponentiating the utterance match scores

$$L = \exp(-a\,\mathrm{d}) \tag{8.9}$$

where *a* is a positive constant (equivalently, the scores are assumed to be proportional to log likelihoods). Likelihood ratios can then be formed using global speaker models or cohorts to normalize *L.*

The template model and its corresponding distance measure is perhaps the most intuitive method. The template method can be dependent or independent of time. An example of a time-independent template model is VQ modeling [45]. All temporal variation is ignored in this model and global averages (e.g., centroids) are all that is used. A time-dependent template model is more complicated because it must accommodate human speaking rate variability.

Template Models

The simplest template model consists of a single template $\overline{\mathbf{x}}$, which is the model for a frame of speech. The match score between the template $\overline{\mathbf{x}}$ for the claimed speaker and an input feature vector $\mathbf{x}_i$ from the unknown user is given by $d\left(\mathbf{x}_i, \overline{\mathbf{x}}\right)$. The model for the claimed speaker could be the centroid (mean) of a set of N training vectors

$$\overline{\mathbf{x}} = \bigcirc\!\!= \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \qquad (8.10)$$

Many different distance measures between the vectors $\mathbf{x}_i$ and $\overline{\mathbf{x}}$ can be expressed as

$$d\left(\mathbf{x}_i, \overline{\mathbf{x}}\right) = (\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}} \mathbf{W} (\mathbf{x}_i - \overline{\mathbf{x}}) \qquad (8.11)$$

where **W** is a weighting matrix. If **W** is an identity matrix, the distance is *Euclidean;* if **W** is the inverse covariance matrix corresponding to mean $\overline{\mathbf{x}}$, then this is the *Mahalanobis distance*. The Mahalanobis distance gives less weight to the components having more variance and is equivalent to a Euclidean distance on principal components, which are the eigenvectors of the original space as determined from the covariance matrix [10].

Dynamic Time Warping

The most popular method to compensate for speaking-rate variability in template-based systems is known as DTW [42]. A text-dependent template model is a sequence of templates $(\overline{\mathbf{x}}_1, \ldots, \overline{\mathbf{x}}_N)$ that must be matched to an input sequence $(\mathbf{x}_1, \ldots, \mathbf{x}_M)$. In general, *N* is not equal to *M* because of timing inconsistencies in human speech. The asymmetric match score *z* is given by

$$z = \sum_{i=1}^{M} \mathrm{d}(\mathbf{x}_i, \overline{\mathbf{x}}_{j(i)}) \qquad (8.12)$$

where the template indices *j(i)* are typically given by a DTW algorithm. Given reference and input signals, the DTW algorithm does a constrained, piecewise linear mapping of one (or both) time axis(es) to align the two signals while minimizing *z*. At the end of the time warping, the accumulated distance is the basis of the match score. This method accounts for the variation over time (trajectories) of parameters corresponding to the dynamic configuration of the articulators and vocal tract. Figure 8.7 shows a warp path for two speech signals using their energies as warp features.

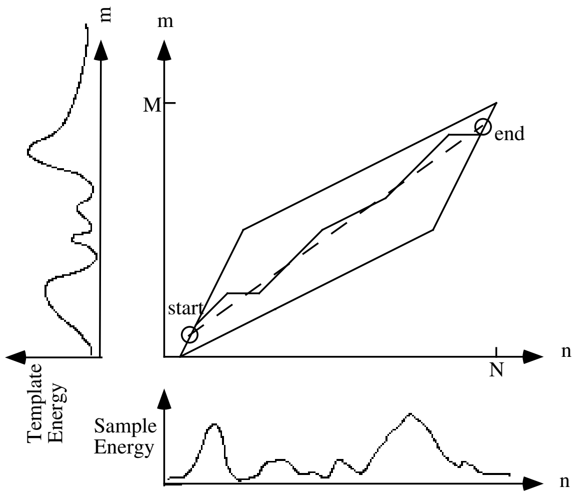

Figure 8.7 DTW path for two energy signals.

If the warp signals were identical, the warp path would be a diagonal line and the warping would have no effect. The Euclidean distance between the two signals in the energy domain is the accumulated deviation off the dashed diagonal warp path. The parallelogram surrounding the warp path represents the Sakoe slope constraints of the warp [42], which act as boundary conditions to prevent excessive warping over a given segment.

Vector Quantization Source Modeling

Another form of template model uses multiple templates to represent frames of speech and is referred to as VQ source modeling [45]. A VQ code book is a collection of codewords and it is typically designed by a clustering procedure. A code book is created for each enrolled speaker using his training data, usually based upon reading a specific text. A pattern match score can be formed as the distance between an input vector $\mathbf{x}_j$ and the minimum distance codeword $\bar{\mathbf{x}}$ in the claimant's VQ code book C. This match score for L frames of speech is

$$z = \sum_{j=1}^{L} \min_{\bar{\mathbf{x}} \in C} \left\{ d\left(\mathbf{x}_j, \bar{\mathbf{x}}\right) \right\} \tag{8.13}$$

The clustering procedure used to form the code book averages out temporal information from the codewords. Thus, there is no need to perform a time alignment. The lack of time warping greatly simplifies the system; however, it neglects speaker-dependent temporal information that may be present in the prompted phrases.

Nearest Neighbors

A technique combining the strengths of the DTW and VQ methods is called nearest neighbors (NN) [17,20]. Unlike the VQ method, the NN method does not cluster the enrollment training data to form a compact code book. Instead, it keeps all the training data and can, therefore, use temporal information.

As shown in Figure 8.8, the claimant's interframe distance matrix is computed by measuring the distance between test-session frames (the input) and the claimant's stored enrollment-session frames. The NN distance is the minimum distance between a test-session frame and the enrollment frames. The NN distances for all the test-session frames are then averaged to form a match score. Similarly, as shown in the rear planes of Figure 8.8, the test-session frames are also measured against a set of stored reference "cohort" speakers to form match scores. The match scores are then combined to form a likelihood ratio approximation [17].

The NN method is one of the most memory- and compute-intensive speaker-verification algorithms. It is also one of the most powerful methods, as illustrated later in Figure 8.10.

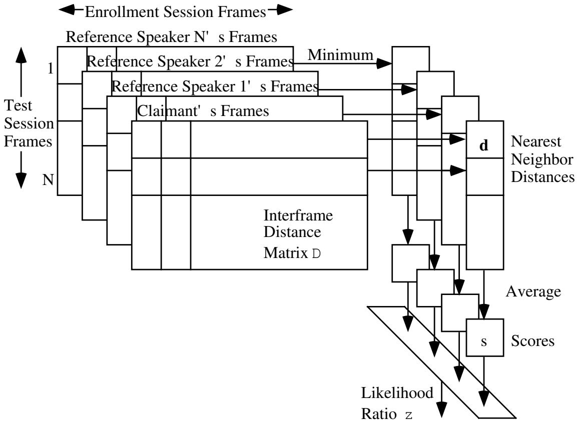
Enrollment Session Frames
Reference Speaker N' s Frames
Reference Speaker 2' s Frames
Minimum
Reference Speaker 1' s Frames
Claimant' s Frames
1
Test Session Frames
N
d
Nearest Neighbor Distances
Interframe Distance Matrix D
Average
s
Scores
Likelihood Ratio z

Figure 8.8 Nearest neighbor method.

Stochastic Models

Template models dominated early work in text-dependent speaker recognition. This deterministic approach is intuitively reasonable, but stochastic models recently have been developed that can offer more flexibility and result in a more theoretically meaningful probabilistic likelihood score.

Using a stochastic model, the pattern-matching problem can be formulated as measuring the likelihood of an observation (a feature vector of a collection of vectors from the unknown speaker) given the speaker model. The observation is a random vector with a conditional probability density function (pdf) that depends upon the speaker. The conditional pdf for the claimed speaker can be estimated from a set of training vectors and, given the estimated density, the probability that the observation was generated by the claimed speaker can be determined.

The estimated pdf can either be a parametric or a nonparametric model. From this model, for each frame of speech (or average of a sequence of frames), the probability that it was generated by the claimed speaker can be estimated. This probability is the match score. If the model is parametric, then a specific pdf is assumed and the appropriate parameters of the density can be estimated using the maximum likelihood estimate. For example, one useful parametric model is the multivariate normal model and it is parameterized by a mean vector $\boldsymbol{\mu}$ and a covariance matrix **C**. In this case, the probability that an observed feature vector $\mathbf{x}_i$ was generated by the model is

$$p(\mathbf{x}_i|\text{model}) = (2\pi)^{-k/2}|\mathbf{C}|^{-1/2}\exp\left\{-\tfrac{1}{2}(\mathbf{x}_i - \mathbf{Q})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{x}_i - \mathbf{Q})\right\} \qquad (8.14)$$

Hence, $p(\mathbf{x}_i|\text{model})$ is the match score. If nothing is known about the true densities, the unknown densities can be approximated by a GMM or nonparametric statistics can be used to find the match score.

The match scores for text-dependent models are given by the probability of a sequence of frames without assuming independence of speech frames. Although a correlation of speech frames is implied by the text-dependent model, deviations of the speech from the model are usually assumed to be independent. This independence assumption enables estimation of utterance likelihoods by multiplying frame likelihoods. The model represents a specific sequence of spoken words.

A stochastic model that is very popular for modeling sequences is the HMM. In conventional Markov models, each state corresponds to a deterministically observable event; thus, the output of such sources in any given state is not random and lacks the flexibility needed here. In an HMM, the observations are a probabilistic function of the state; i.e., the model is a doubly embedded stochastic process where the underlying stochastic process is not directly observable (it is hidden). The HMM can only be viewed through another set of stochastic processes that produce the sequence of observations [32]. The HMM is a finite-state machine, where a pdf (or feature vector stochastic model) $p(\mathbf{x}|s_i)$ is associated with each state s_i (the main underlying model). The states are connected by a transition network, where the state transition probabilities are $a_{ij} = p(s_i|s_j)$. For example, a hypothetical three-state HMM is illustrated in Figure 8.9.

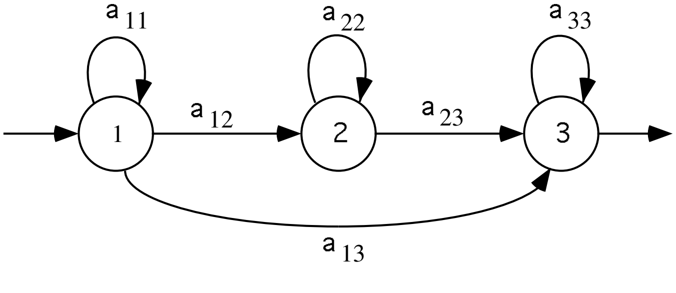

Figure 8.9 An example of a three-state HMM.

The probability that a sequence of speech frames was generated by this model can be found by Baum-Welch decoding [30,31]. This likelihood is the score for L frames of input speech given the model

$$p(\mathbf{x}(1;L)|\text{model}) = \sum_{\substack{\text{all state}\\ \text{sequences}}} \prod_{i=1}^{L} p(\mathbf{x}_i|s_i)\,p(s_i|s_{i-1}) \qquad (8.15)$$

This is a theoretically meaningful score. HMM-based methods have been shown to be comparable in performance to conventional VQ methods in text-independent testing [47] and more recently to outperform conventional methods in text-dependent testing (e.g., [35]).

5. Classification and Decision Theory

Having computed a match score between the input speech-feature vector and a model of the claimed speaker's voice, a verification decision is made whether to accept or reject the speaker or request another utterance (or, without a claimed identity, an identification decision is made). If a verification system accepts an impostor, it makes a false acceptance (FA) error. If the system rejects a valid user, it makes a false rejection (FR) error. The FA and FR errors can be traded off by adjusting the decision threshold, as shown by a Receiver Operating Characteristic (ROC) curve. The operating point where the FA and FR are equal corresponds to the equal error rate.

The accept or reject decision process can be an accept, continue, time-out, or reject hypothesis-testing problem. In this case, the decision making, or classification, procedure is a sequential hypothesis-testing problem [48]. For a brief overview of the decision theory involved, please refer to [6].

6. Performance

Using the YOHO prerecorded speaker-verification database, the following results on wolves and sheep were measured. The impostor testing was simulated by randomly selecting a valid user (a potential wolf) and altering his/her identity claim to match that of a randomly selected target user (a potential sheep). Because the potential wolf is not intentionally attempting to masquerade as the potential sheep, this is referred to as the "casual impostor" paradigm. Testing the system to a certain confidence level implies a minimum requirement for the number of trials. In this testing, there were 9,300 simulated impostor trials to test to the desired confidence [5,17].

DTW System

The DTW ASV system tested here was created by Higgins, *et al.* [18]. This system is a variation on a DTW approach that introduced likelihood ratio scoring via cohort normalization in which the input utterance is compared with the claimant's voice model and with an alternate model composed of models of other users with similar voices. Likelihood ratio scoring allows for a fixed, speaker-independent, phrase-independent acceptance criterion. Pseudorandomized phrase prompting, consistent with the YOHO corpus, is used in combination with speech recognition to reduce the threat of playback (e.g., tape recorder) attacks. The enrollment algorithm creates users' voice models based upon subword models (e.g., "twen," "ti," and "six"). Enrollment begins with a generic male or female template for each subword and results in a speaker-specific template model for each subword. These models and their estimated word endpoints are successively refined by including more examples collected from the enrollment speech material [18].

Cross-speaker testing (casual impostors) was performed, confusion matrices for each system were generated, wolves and sheep of DTW and NN systems were identified, and errors were analyzed.

Table 8.5 shows two measures of wolves and sheep for the DTW system: those who were wolves or sheep at least once and those who were wolves or sheep at least twice. Thus, FA errors occur in a very narrow portion of the 186-person population, especially if two errors are required to designate a person as a wolf or sheep. The difficulty in acquiring enough data to adequately represent the wolf and sheep populations makes it challenging to study these errors.

186 Subjects of the YOHO Database	
At least one FA Error	At least two FA Errors
17 Wolves (9%)	2 Wolves (1%)
11 Sheep (6%)	5 Sheep (3%)

Table 8.5 Known wolves and sheep of the DTW system.

The DTW system made 19 FA errors over the 9,300 impostor trials. Table 8.6 shows that these 19 pairs of wolves and sheep have interesting characteristics. The database contains four times as many males as it does females, but the 18:1 ratio of male wolves to female wolves is disproportionate. It is also interesting to note that one male wolf successfully preyed upon three different female sheep. The YOHO corpus provides at least 19 pairs of wolves and sheep under the DTW ASV system for further investigation.

19 FA errors across 9300 impostor trials		
Number of FA errors	Wolf sex	Sheep sex
15	Males	Males
1	Female	Female
3	1 Male	3 Females

Table 8.6 Wolf and sheep distribution by sex.

ROC of DTW and NN Systems

Figure 8.10 shows the NN system's ROC curve and a point on the ROC for the DTW system (ROCs of better systems are closer to the origin). The NN system was the first one known to meet the 0.1% FA and 1% FR performance level at the 80% confidence level and it outperforms the DTW system by about half an order of magnitude.

These overall error rates do not show the individual wolf and sheep populations of the two systems. As shown in Figures 8.11-8.14, the two systems commit different errors.

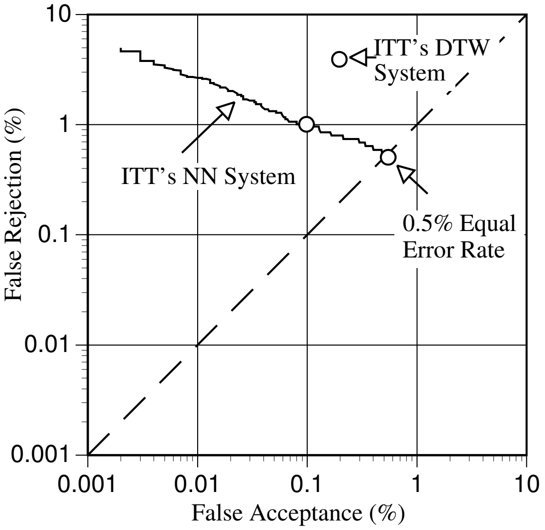

Figure 8.10 Receiver operating characteristics.

Wolves and Sheep

FA errors due to individual wolves and sheep are shown in the 3-D histogram plots of Figures 8.11 through 8.14. Figure 8.11 shows the individual speakers who were falsely accepted as other speakers by the DTW system. For example, the person with an identification number of 97328 is never a wolf and is a sheep once under the DTW system.

The DTW system rarely has the same speaker as both a wolf and a sheep (there are only two exceptions in this data). These exceptions, called *wolf-sheep*, probably have poor models because they match a sheep's model more closely than their own and a wolf's model also matches their model more closely than their own. These *wolf-sheep* would likely benefit from retraining to improve their models.

Now let us look at the NN system. Figure 8.12 shows the FA errors committed by the NN system. Two speakers, who are sheep, are seen to dominate the NN system's FA errors. A dramatic performance improvement would result if these two speakers were recognized correctly by the system.

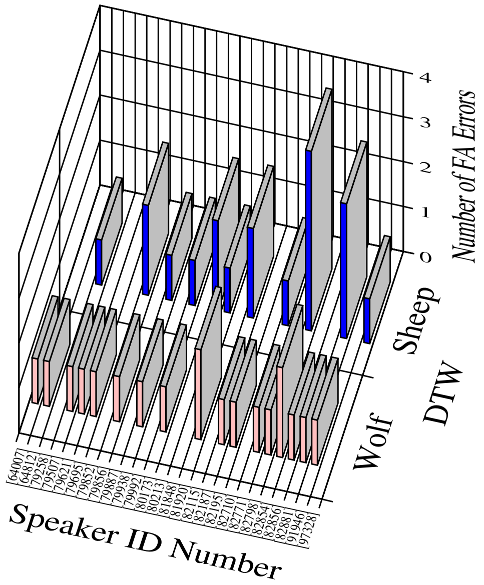

Figure 8.11 Speaker versus FA errors for the DTW system's wolves and sheep.

Now we will investigate the relations between the NN and DTW systems. Figure 8.13 shows the sheep of the NN and DTW systems. The two sheep that dominate the FA errors of the NN system are shown not to be sheep in the DTW system. This suggests the potential for making a significant performance improvement by combining the systems.

Figure 8.14 shows that the wolves of the NN system are dominated by a few individuals who do not cause errors in the DTW system. Again, this suggests the potential for realizing a performance improvement by combining elements of the NN and DTW systems. Along these lines, a high-performance speaker detection system consisting of eight combined systems has been demonstrated recently [27].

7. Conclusions

Automatic speaker recognition is the use of a machine to recognize a person from a spoken phrase. Speaker-recognition systems can be used to *identify* a particular person or to *verify* a person's claimed identity. Speech processing, speech production, and features and pattern matching for speaker recognition were introduced. Recognition accuracy was shown by coarse-grain ROC curves and fine-grain histograms revealed the wolves and sheep of two example systems. Speaker recognition systems can achieve 0.5% equal error rates at the 80% confidence level in the benign real-world conditions considered here.

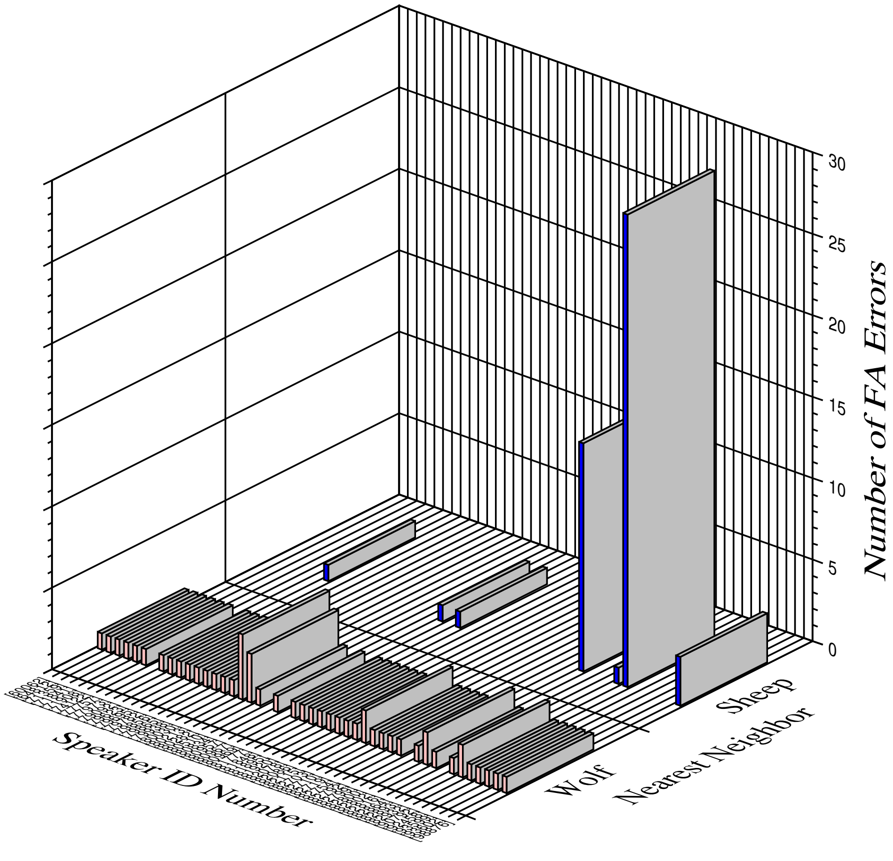

Figure 8.12 Speaker versus FA errors for NN system's wolves and sheep.

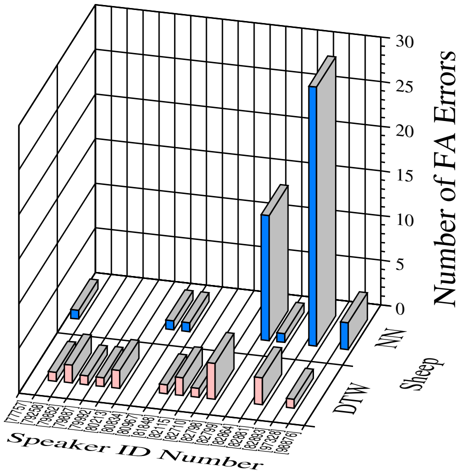

Figure 8.13 Speaker versus FA errors for DTW and NN systems' sheep.

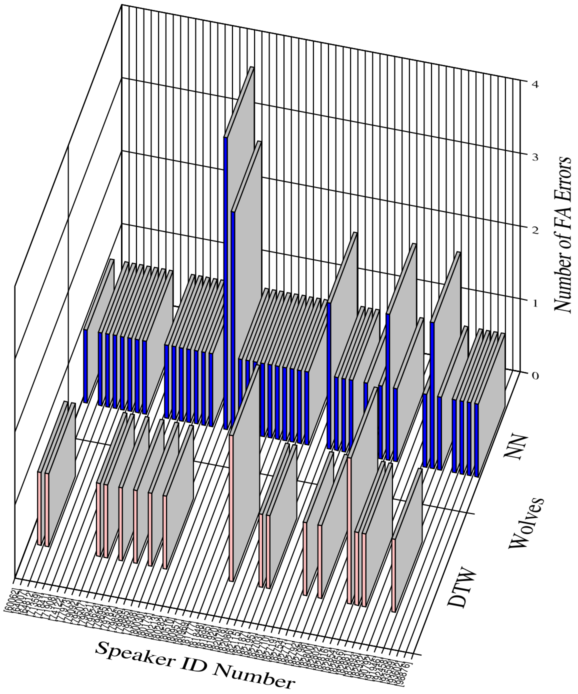

Figure 8.14 Speaker versus FA errors for DTW and NN systems' wolves.

References

[1] B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," *Journal of the Acoustical Society of America,* Vol. 55, No. 6, pp. 1304-1312, 1974.

[2] B. S. Atal, "Automatic Recognition of Speakers from Their Voices," *Proceedings of the IEEE* Vol. 64, pp. 460-475, 1976.

[3] J. Attili, M. Savic, and J. Campbell. "A TMS32020-Based Real Time, Text-Independent, Automatic Speaker Verification System," In *International Conference on Acoustics, Speech, and Signal Processing in New York,* IEEE, pp. 599-602, 1988.

[4] J. P. Campbell, T. E. Tremain, and V. C. Welch. "The Federal Standard 1016 4800 bps CELP Voice Coder," *Digital Signal Processing*, Vol. 1, No. 3, pp. 145 -155, 1991.

[5] J. P. Campbell, "Testing with The YOHO CD-ROM Voice Verification Corpus," In *International Conference on Acoustics, Speech, and Signal Processing in Detroit,* IEEE, pp. 341-344, 1995. Available: http://www.biometrics.org/.

[6] J. P. Campbell, "Speaker Recognition: A Tutorial." *Proceedings of the IEEE,* Vol. 85, No. 9, pp. 1437-1462, 1997.

[7] C. Che and Q. Lin, "Speaker recognition using HMM with experiments on the YOHO database," In *EUROSPEECH in Madrid*, ESCA, pp. 625- 628, 1995.

[8] J. Colombi, D. Ruck, S. Rogers, M. Oxley, and T. Anderson, "Cohort Selection and Word Grammer Effects for Speaker Recognition," In *International Conference on Acoustics, Speech, and Signal Processing in Atlanta,* IEEE, pp. 85- 88, 1996.

[9] G. R. Doddington, "Speaker Recognition—Identifying People by their Voices," *Proceedings of the IEEE*, Vol. 73, No. 11, pp. 1651-1664, 1985.

[10] R. Duda and P. Hart. *Pattern Classification and Scene Analysis.* New York: Wiley, 1973.

[11] J. Flanagan, *Speech Analysis Synthesis and Perception.* 2nd ed., Berlin: Springer-Verlag, 1972.

[12] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-29, No. 2 , pp. 254-272, 1981.

[13] S. Furui, "Speaker-Dependent-Feature Extraction, Recognition and Processing Techniques." *Speech Communication*,Vol. 10, pp. 505-520, 1991.

[14] S. Furui, "Recent Advances in Speaker Recognition," *Pattern Recognition Letters*, Vol. 18, pp. 859-872, 1997.

[15] H. Gish and M. Schmidt, "Text-Independent Speaker Identification," *IEEE Signal Processing Magazine*, Vol. 11, No. 4, pp. 18-32, 1994.

[16] R. Gnanadesikan and J. R. Kettenring. "Discriminant Analysis and Clustering," *Statistical Science*, Vol. 4, No. 1, pp. 34-69, 1989.

[17] A. Higgins, "YOHO Speaker Verification," Speech Research Symposium, Baltimore, 1990.

[18] A. Higgins, L. Bahler, and J. Porter, "Speaker Verification Using Randomized Phrase Prompting," *Digital Signal Processing*, Vol. 1, No. 2 , pp. 89-106, 1991.

[19] A. L. Higgins and R. E. Wohlford, "A New Method of Text-Independent Speaker Recognition," In *International Conference on Acoustics, Speech, and Signal Processing in Tokyo,* IEEE, pp. 869-872, 1986.

[20] A. Higgins, L. Bahler, and J. Porter, "Voice Identification Using Nearest Neighbor Distance Measure," In *International Conference on Acoustics, Speech, and Signal Processing in Minneapolis,* IEEE, pp. 375-378, 1993.

[21] F. Itakura, "Line Spectrum Representation of Linear Predictive Coefficients," *Transactions of the Committee on Speech Research, Acoustical Society of Japan*, Vol. S75, No. 34, 1975.

[22] G. Kang and L. Fransen, *Low Bit Rate Speech Encoder Based on Line-Spectrum-Frequency,* NRL, NRL Report 8857, 1985.

[23] K. P. Li and E. H. Wrench, "Text-Independent Speaker Recognition with Short Utterances," In *International Conference on Acoustics, Speech, and Signal Processing in Boston,* IEEE, pp. 555-558, 1983.

[24] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proceedings of the IEEE*, Vol. 63, pp. 561-580, 1975.

[25] R. Mammone, X. Zhang, and R. Ramachandran, "Robust Speaker Recognition-A Feature-based Approach," *IEEE Signal Processing Magazine*, Vol. 13, No. 5, pp. 58-71, 1996.

[26] J. D. Markel and S. B. Davis, "Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base," *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-27, No. 1, pp. 74-82, 1979.

[27] A. Martin and M. Przybocki, "1997 Speaker Recognition Evaluation," In *Speaker Recognition Workshop,* editor A. Martin (NIST). Section 2. Maritime Institute of Technology, Linthicum Heights, Maryland, June, pp. 25-26, 1997. Available: ftp://jaguar.ncsl.nist.gov/speaker/ and http://www.nist.gov/itl/div894/894.01/.

[28] D. O'Shaughnessy, *Speech Communication, Human and Machine.* Digital Signal Processing, Reading: Addison-Wesley, 1987.

[29] G. Papcun, "Commensurability Among Biometric Systems: How to Know When Three Apples Probably Equals Seven Oranges," In *Proceedings of the Biometric Consortium, 9th*

Meeting, editor J. Campbell (NSA). Holiday Inn, Crystal City, Virginia, April, 8-9, 1997. Available: http://www.biometrics.org/.

[30] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257-286, 1989.

[31] L. Rabiner and B. H. Juang, "An Introduction to Hidden Markov Models," *IEEE ASSP Magazine*, Vol. 3, pp. 4-16, January 1986.

[32] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Signal Processing, editor A. Oppenheim. Englewood Cliffs: Prentice-Hall, 1993.

[33] L. Rabiner and R. Schafer. *Digital Processing of Speech Signals*, Signal Processing, editor A. Oppenheim. Englewood Cliffs: Prentice-Hall, 1978.

[34] D. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models," *Speech Communication*, Vol. 17, pp. 91-108, 1995.

[35] D. Reynolds and B. Carlson, "Text-Dependent Speaker Verification Using Decoupled and Integrated Speaker and Speech Recognizers," In *EUROSPEECH in Madrid*, ESCA, pp. 647-650, 1995.

[36] D. Reynolds and R. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 1, pp. 72-83, 1995.

[37] D. Reynolds, "M.I.T. Lincoln Laboratory Site Presentation," In *Speaker Recognition Workshop,* editor A. Martin (NIST). Section 5. Maritime Institute of Technology, Linthicum Heights, Maryland, March, pp. 27-28, 1996. Available: ftp://jaguar.ncsl.nist.gov/speaker/ and http://www.nist.gov/itl/div894/894.01/.

[38] A. Rosenberg, "Automatic Speaker Verification: A Review," *Proceedings of the IEEE*, Vol. 64, No. 4, pp. 475-487, 1976.

[39] E. Rosenberg and F. K. Soong, "Recent Research in Automatic Speaker Recognition," In *Advances in Speech Signal Processing,* ed. S. Furui and M. M. Sondhi. pp. 701-738. New York: Marcel Dekker, 1992.

[40] E. Rosenberg, J. DeLong, C-H. Lee, B-H. Juang, and F. K. Soong, "The Use of Cohort Normalized Scores for Speaker Verification," In *International Conference on Spoken Language Processing in Banff,* University of Alberta, pp. 599-602, 1992.

[41] S. Saito and K. Nakata. *Fundamentals of Speech Signal Processing.* Tokyo: Academic Press, 1985.

[42] H. Sakoe and S. Chiba. "Dynamic Programming Algorithm Optimization for Spoken Word Recognition." *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-26, No. 1, pp. 43-49, 1978.

[43] R. Schwartz, S. Roucos, and M. Berouti, "The Application of Probability Density Estimation to Text Independent Speaker Identification," In *International Conference on Acoustics, Speech, and Signal Processing in Paris,* IEEE, pp. 1649-1652, 1982.

[44] F. Soong, A. Rosenberg, L. Rabiner, and B-H. Juang, "A Vector Quantization Approach to Speaker Recognition," In *International Conference on Acoustics, Speech, and Signal Processing in Florida,* IEEE, pp. 387-390, 1985.

[45] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang, "A Vector Quantization Approach to Speaker Recognition." *AT&T Technical Journal*, Vol. 66, No. 2 , pp. 14-26, 1987.

[46] A. Sutherland and M. Jack, "Speaker Verification." In *Aspects of Speech Technology,* editors M. Jack and J. Laver, Edinburgh: Edinburgh University Press, pp. 185-215, 1988.

[47] N. Z. Tishby, "On the Application of Mixture AR Hidden Markov Models to Text Independent Speaker Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 39, No. 3, pp. 563 – 570, 1991.

[48] A. Wald, *Sequential Analysis.* New York: Wiley, 1947.