SPEAKER RECOGNITION

The focus of this article is on facilities and network access-control applications of speaker recognition. Speech processing is a diverse field with many applications. Figure 1 shows a few of these areas and how speaker recognition relates to the rest of the field.

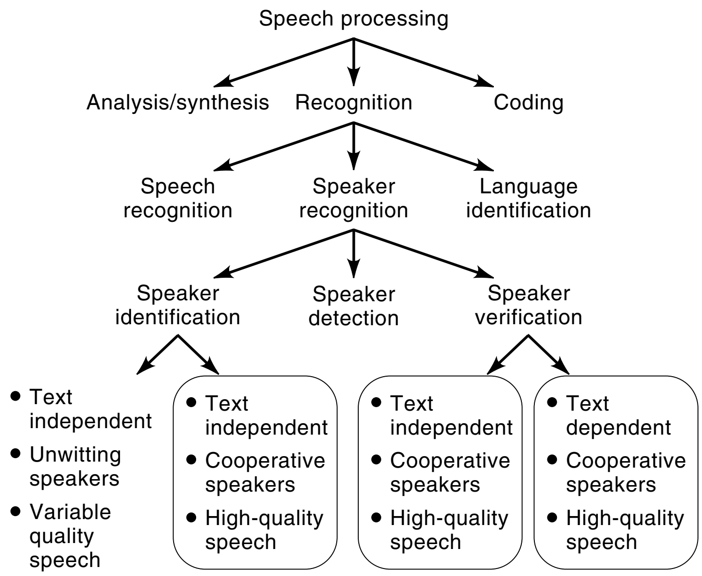

Figure 1. Speech processing.

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 or her 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) whether the person is unknown. General overviews of speaker recognition have been given by Atal (1), Doddington (2), Furui (3), O'Shaughnessy (4), Rosenberg (5), Rosenberg and Soong (6), and Sutherland and Jack (7).

Speaker verification is defined as deciding if a speaker is who he or she 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 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 Fig. 2. The claimant, who has previously enrolled in the system, presents an encrypted smart card containing identification information. The claimant then attempts to be authenticated by speaking a prompted phrase(s) into the microphone. There is generally a tradeoff between accuracy and test-session duration. In addition to the voice itself, ambient room noise and delayed versions of the 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 (possi-

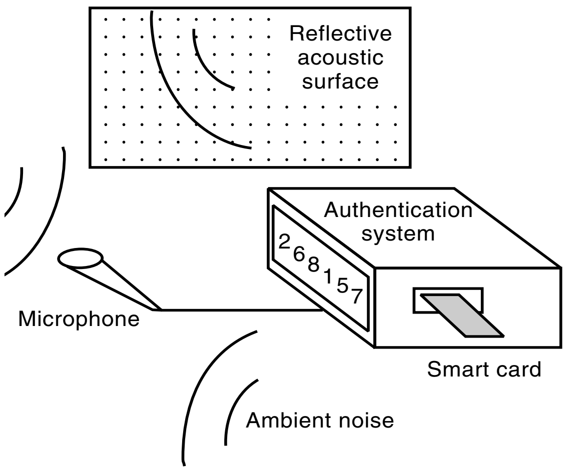

Figure 2. Typical speaker verification setup.

bly on a smart card) for use in later verification sessions. There is also generally a tradeoff between accuracy and the duration and number of enrollment sessions.

Many factors can contribute to verification and identification errors. Table 1 lists some of the human and environmental factors that contribute to these errors, a few of which are shown in Fig. 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.

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 software.

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 (8,9), ordinary human changes (e.g., time-of-day voice changes and minor head colds), and mimicry by humans and tape recorders (10).

Table 1. Sources of Verification Error

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)

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 among speakers.

Generic Speaker Verification

The general approach to ASV consists of five steps: digital speech data acquisition, feature extraction, pattern matching, decision acceptance/rejection, and enrollment to generate speaker reference models. A block diagram of this procedure is shown in Fig. 3. Feature extraction maps each interval of speech to a multidimensional feature space. (A speech interval typically spans 10 ms 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 $\mathbf{z}_i$ 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 either to accept or to reject the claimant according to the match score or sequence of match scores, which is a hypothesis-testing problem.

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.

OVERVIEW

The purpose of these introductory remarks is to present a general framework and motivation for speaker recognition, an overview of the entire article, and a presentation of previous work in speaker recognition.

The section entitled "Speech Processing" contains an overview of speech processing, including speech signal acquisition, the database used in later experiments, speech production, linear prediction (LP), transformations, and the cepstrum. The section entitled "Feature Selection and Measures" presents feature selection, the divergence measure, and the Bhattacharyya distance. This section is highlighted by the development of the divergence shape measure and the Bhattacharyya distance shape. The next section introduces pattern matching. It is followed by a section that presents classification, decision theory, and receiver operating characteristic (ROC) curves. The section entitled "A New Speaker Recognition System" describes a simple, but effective, speaker-recognition algorithm. The section entitled "Performance" demonstrates the performance of various speaker-recognition algorithms, and the last section summarizes this article.

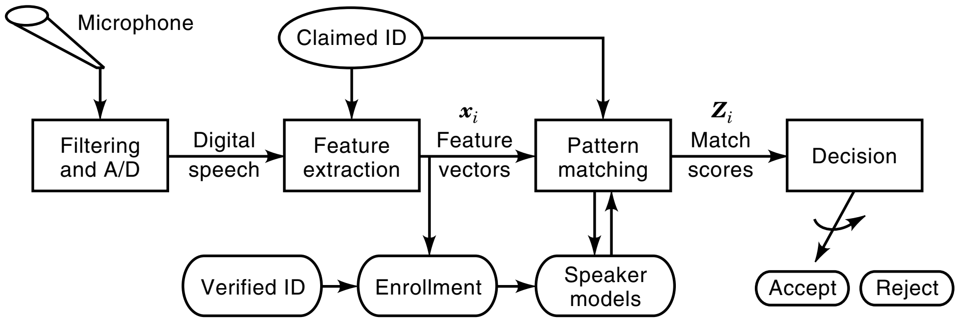

Figure 3. Generic speaker verification system.

PREVIOUS WORK

There is considerable speaker recognition activity in industry, national laboratories, and universities. Among the institutions that have researched and designed several generations of speaker-recognition systems are AT&T (and its derivatives); Bolt, Beranek and Newman (BBN); the Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP, Switzerland); ITT Industries (ITT); Massachusetts Institute of Technology Lincoln Laboratory (MIT-LL); National Tsing Hua University (Taiwan); Nagoya University (Japan); Nippon Telegraph and Telephone (NTT, Japan); Rensselaer Polytechnic Institute (RPI); Rutgers University; and Texas Instruments (TI). The majority of ASV research is directed at verification over telephone lines. Sandia National Laboratories, the National Institute of Standards and Technology (11), and the National Security Agency (12) have conducted evaluations of speaker-recognition systems.

Table 2 shows a sampling of the chronological advancement in speaker verification. The following terms are used to define the columns in Table 2: "Source" refers to a citation in the references (13–26), "Org" is the company or school where the work was done, "Features" are the signal measurements (e.g., cepstrum), "Method" is the heart of the pattern-matching process, "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, "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. These data are 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 (27), differences in partitioning the impostor and cohort sets, and differences in known versus unknown impostors (12)]. 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 (28) and Reynold's Gaussian Mixture Model (9), 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 (2,29).

The general trend shows accuracy improvement over time with larger tests (enabled by larger databases), thus increasing our confidence in the performance measurements. For high-security applications, these speaker recognition systems would need to be used in combination with other authenticators (e.g., smart card). The performance of current speaker recognition systems, however, makes them suitable for many practical applications. 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 United States steal $3 to $5 billion annually. Automatic speaker recognition technology could substantially reduce this crime by reducing these fraudulent transactions.

As automatic-speaker verification systems gain widespread use, it is imperative to understand the errors made by these systems. There are two types of errors: the false acceptance of an invalid user (FA or Type I) and the false rejection of a valid user (FR or type II). 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.

After reviewing the methods of speaker recognition, a simple speaker recognition system will be presented. A database of 186 people collected over a 3 month period was used in closed-set speaker identification experiments. A speaker recognition system using methods presented here is practical to implement in software on a modest personal computer. The example system uses features and measures for speaker recognition based upon speaker discrimination criterion (the ultimate goal of any recognition system). Experimental results show that these new features and measures yield 1.1% closed-set speaker identification error on databases of 44 and 43 people. The features and measures use long-term statistics based upon an information-theoretic shape measure between line spectrum pair (LSP) frequency features. This new measure, the *divergence shape,* can be interpreted geometrically as the shape of an information-theoretic measure called divergence. The LSPs were found to be very effective features in this divergence shape measure.

Table 2. Selected Chronology of Speaker Recognition Progress

Source	Org	Features	Method	Input	Text	Pop	Error
Atal (13)	AT&T	Cepstrum	Pattern match	Lab	Dependent	10	i: 2%@0.5s v: 2%@1s
Markel and Davis (14)	STI	LP	Long-term statistics	Lab	Independent	17	i: 2%@39s
Furui (15)	AT&T	Normalized cepstrum	Pattern match	Telephone	Dependent	10	v: 0.2%@3s
Schwartz et al. (16)	BBN	LAR	Nonparametric pdf	Telephone	Independent	21	i: 2.5%@2s
Li and Wrench (17)	ITT	LP, cepstrum	Pattern match	Lab	Independent	11	i: 21%@3s i: 4%@10s
Doddington (2)	TI	Filter-bank	DTW	Lab	Dependent	200	v: 0.8%@6s
Soong et al. (18)	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. (20)	RPI	Cepstrum, LP, autocorr.	Projected long-term statistics	Lab	Dependent	90	v: 1%@3s
Higgins et al. (10)	ITT	LAR, LP-cepstrum	DTW likelihood scoring	Office	Dependent	186	v: 1.7%@10s
Tishby (21)	AT&T	LP	HMM (AR mix)	Telephone	10 isolated digits	100	v: 2.8%@1.5s v: 0.8%@3.5s
Reynolds (22); Reynolds and Carlson (23)	MIT-LL	Mel-cepstrum	HMM (GMM)	Office	Dependent	138	i: 0.8%@10s v: 0.12%@10s
Che and Lin (24)	Rutgers	Cepstrum	HMM	Office	Dependent	138	i: 0.56%@2.5s i: 0.14%@10s v: 0.62%@2.5s
Colombi et al. (25)	AFIT	Cep, eng d cep, dd cep	HMM monophone	Office	Dependent	138	i: 0.22%@10s v: 0.28%@10s
Reynolds (26)	MIT-LL	Mel-cepstrum, mel-d cepstrum	HMM (GMM)	Telephone	Independent	416	v: 11%/16%@3s v: 6%/8%@10s v: 3%/5%@30s matched/mismatched handset

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

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 converter (ADC). Today's ADCs for speech applications typically 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 long-distance 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 3 describes the YOHO database (30). YOHO is available from the Linguistic Data Consortium (University of Pennsylvania) and test plans have been developed for its use (12). 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 Fig. 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 US 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).

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 to be the speech production organs above the vocal folds. As shown in Fig. 4 (31), this includes the 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 (31).

The vocal folds (formerly known as vocal cords) are shown in Fig. 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 the 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.

Table 3. The YOHO Corpus

"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 Gb of data

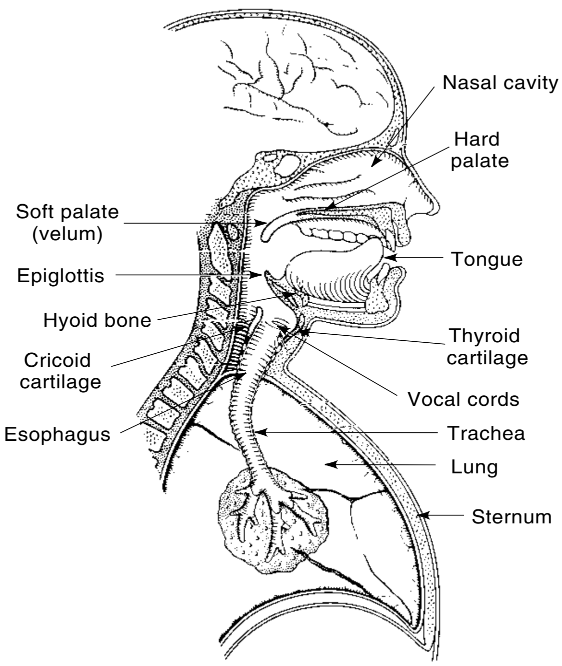

Figure 4. Human vocal system. Reprinted with permission from Springer-Verlag (31).

Voice verification systems typically use features derived only from the vocal tract. As seen in Fig. 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.

Phonated excitation (phonation) occurs when airflow is modulated by the vocal folds. When the vocal folds are closed, pressure builds up underneath them until they blow apart. Then the folds are drawn back together again by tension, elasticity, and the Bernoulli effect. This pulsed air stream, arising from the oscillating vocal folds, excites the vocal tract. The frequency of oscillation is called the fundamental frequency, and it depends upon the length, tension, and mass of the vocal folds. Thus, fundamental frequency is another distinguishing characteristic that is physically based.

Whispered excitation is produced by airflow rushing through a small triangular opening between the arytenoid cartilages at the rear of the nearly closed vocal folds. This results in turbulent airflow, which has a wideband noise characteristic (32).

Frication excitation is produced by constrictions in the vocal tract. The place, shape, and degree of constriction determine the shape of the broadband noise excitation. As the constriction moves forward, the spectral concentration generally increases in frequency. Sounds generated by frication are called *fricatives* or *sibilants.* Frication can occur without phonation (e.g., "s" as in sass) or with phonation (e.g., "z" as in zoos).

Compression excitation results from releasing a completely closed and pressurized vocal tract. This results in silence (during pressure accumulation) followed by a short noise burst. If the release is sudden, a *stop* or *plosive* is generated. If the release is gradual, an *affricate* is formed.

Vibration excitation is caused by air being forced through a closure other than the vocal folds, especially at the tongue (e.g., trilled "r").

Speech produced by phonated excitation is called *voiced,* speech produced by phonated excitation plus frication is called *mixed voiced,* and speech produced by other types of excitation is called *unvoiced.* Because of the differences in the manner of production, it's reasonable to expect some speech models to be more accurate for certain classes of excitation than others. Unlike phonation and whispering, the places of frication, compression, and vibration excitation are actually inside the vocal tract itself. This could cause difficulties for models that assume an excitation at the bottom end of the vocal tract. For example, the linear prediction model assumes a vocal tract excited at a closed end. Phonation excitation is the only one that approximates this assumption. Thus, it is reasonable to use different models or different weighting for those regions of speech that violate any modeling assumptions.

The respiratory (thoracic area) plays a role in the resonance properties of the vocal system. The trachea is a pipe, typically 12 cm long and 2 cm in diameter, made up of rings of cartilage joined by connective tissue joining the lungs and the larynx. When the vocal folds vibrate, there are resonances above and below the folds. Subglottal resonances are largely dependent upon the properties of the trachea (33). Because of this physiological dependence, subglottal resonances have speaker-dependent properties.

Other physiological speaker-dependent properties include vital capacity (the maximum volume of air one can blow out after maximum intake), maximum phonation time (the maximum duration a syllable can be sustained), phonation quotient (ratio of vital capacity to maximum phonation time), and glottal airflow (amount of air going through vocal folds). Because sound and airflow are different, these dimensions can be difficult to acquire from the acoustic signal alone; however, Plumpe (34) has shown encouraging speaker identification research using the glottal flow derivative waveform estimated from the acoustic signal.

Other aspects of speech production that could be useful for discriminating between speakers are learned characteristics, including speaking rate, prosodic effects, and dialect (which might be captured spectrally as a systematic shift in formant frequencies).

Linear Prediction

The all-pole LP models a signal s_n by a linear combination of its past values and a scaled present input (35)

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

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

$$\hat{s}_n = -\sum_{k=1}^{p} a_k \cdot s_{n-k} \tag{2}$$

This greatly simplifies the problem of estimating the a_k because the source (i.e., the glottal input) and filter (i.e., the vocal tract) have been decoupled. 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

$$e_n = s_n - \hat{s}_n = s_n + \sum_{k=1}^{p} a_k \cdot s_{n-k} \tag{3}$$

Therefore, the prediction error e_n is identical to the scaled input signal $G \cdot u_n$. Letting E represent the mean squared error (MSE),

$$E = \sum_n e_n^2 = \sum_n \left[s_n + \sum_{k=1}^{p} a_k \cdot s_{n-k} \right]^2 \tag{4}$$

The minimum MSE criteria resulting from

$$\frac{\partial E}{\partial a_i} = 0, \vee i = 1, 2, \ldots, p \tag{5}$$

is

$$\sum_{k=1}^{p} a_k \cdot \sum_n s_{n-k} s_{n-i} = -\sum_n s_n s_{n-i} \vee i = 1, 2, \ldots, p \tag{6}$$

where the summation ranges on n have been intentionally omitted for generality. If the summation is of infinite extent (or over the nonzero length of a finite extent window) (36), the summations on s are the autocorrelations at lags $i - k$ for the left sum and at lag i for the right sum. This results in the "autocorrelation method" of LP analysis. (Other LP methods, such as covariance and Burg's, arise from variations on windowing, the extent of the signal, and whether the summations on s are one or two sided.) The time-averaged estimates of the autocorrelation at lag τ can be expressed as

$$R_\tau = \sum_{i=0}^{N-1-\tau} s(i) \cdot s(i+\tau) \tag{7}$$

The autocorrelation method yields the system of equations named after Yule's pioneering all-pole modeling in sunspot analysis and given by Eq. (8).

$$\begin{bmatrix} R_0 & R_1 & R_2 & \cdots & R_{p-1} \\ R_1 & R_0 & R_1 & \ddots & R_{p-2} \\ R_2 & R_1 & R_0 & \ddots & R_{p-3} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ R_{p-1} & R_{p-2} & R_{p-3} & \cdots & R_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ \vdots \\ R_p \end{bmatrix} \tag{8}$$

The LP model parameters we seek are a_k. For a pth-order prediction, the speech signal is modeled by a p-dimensional a_k vector. As the Yule–Walker equation shows, this requires the computation of $p + 1$ autocorrelations and matrix inversion. The matrix inversion problem is greatly simplified because of the symmetric Toeplitz autocorrelation matrix on the left-hand side of Eq. (8), $\mathbf{R} = R_{|i-j|}$, and the form of the autocorrelation vector on the right, which are exploited by Durbin's recursive algorithm. This algorithm is the most efficient method known for solving this particular system of equations (35). Note that in the process of solving for the predictor coefficients a_k of order p, the a_k for all orders less than p are obtained with their corresponding mean-square prediction error: $\mathrm{MSE}_i = E_i/R_0$. In each recursion of Durbin's algorithm, the prediction order is increased and the corresponding error is determined; this can be monitored as a stopping criterion on the prediction order p.

$$\begin{aligned}
&E_0 = R_0 \\
&\left.\begin{aligned}
&k_i = -\frac{R_i + \sum_{j=1}^{i-1} a_j^{(i-1)} R_{i-j}}{E_{i-1}} && \vee 1 \le i \le p \\
&a_i^{(i)} = k_1 \\
&a_j^{(j)} = a_j^{(i-1)} + k_i a_{t-j}^{(i-1)} && \vee 1 \le j \le i-1 \\
&E_i = (1 - k_i^2) E_{i-1}
\end{aligned}\right\} \vee i = 1, 2, \ldots, p \\
&a_j = a_j^{(p)} \qquad \vee\, 1 \le j \le p
\end{aligned} \tag{9}$$

Using the a_k model parameters, Eq. (10) 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 PCs.

$$s_n = -\sum_{k=1}^{p} a_k \cdot s_{n-k} + e_n \tag{10}$$

From Eq. (1), the LP transfer function is defined as

$$H(z) \equiv \frac{S(z)}{U(z)} \equiv \frac{Z[s_n]}{Z[u_n]} \tag{11}$$

which yields

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} \equiv \frac{G}{A(z)} \tag{12}$$

where $A(z)$ is known as the pth-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 MSE is minimized because it allows a simple, closed-form solution of the PCs. Minimizing MSE error tends to produce a flat (band-limited white) magnitude spectrum of the error signal. Hence, the inverse filter $A(z)$ is also known as a "whitening" filter.

If a voiced speech signal fits the model, then the residual is an impulse train that repeats at the rate of vocal-fold vibration. Therefore, the maximum prediction errors (residual peaks) occur at the vocal-fold vibration rate. (Many "pitch detection" algorithms exploit this property.) Thus, in the time domain, the majority of energy lost in the PCs occurs in the vicinity of these "pitch peaks."

Features are constructed from the speech model parameters [e.g., the a_k shown in Eq. (12)]. 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; LSP frequencies, introduced by Itakura (37–39); and the LP cepstrum (40).

Reflection Coefficients. If Durbin's algorithm 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 (40)

$$\left.\begin{aligned}
&\alpha_j^{(p)} = a_j \\
&k_i = \alpha_i^{(i)} \\
&\alpha_j^{(i-1)} = \frac{\alpha_j^{(i)} + \alpha_i^{(i)} \cdot \alpha_{i-j}^{(i)}}{1 - k_i^2} \vee 1 \le j \le i-1
\end{aligned}\right\} \vee i = p, p-1, \ldots, 1 \tag{13}$$

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 $A_0, A_1, \ldots, A_5$ could look like Fig. 5. This series of five tubes represents a fourth-order system that might fit a vocal tract minus the nasal cavity. Given boundary conditions, the reflection coefficients are determined by the ratios of the adjacent cross-sectional areas (40). For a pth-order system, the boundary conditions given in Eq. (14) correspond to a closed glottis (zero area) and a large area following the lips

$$\begin{aligned}
&A_0 = 0 \\
&A_{p+1} \gg A_p \\
&k_i = \frac{A_{i+1} - A_i}{A_{i+1} + A_i} \vee i = 1, 2, \ldots, p
\end{aligned} \tag{14}$$

Thus, the reflection coefficients can be derived from an acoustic tube model or an autoregressive model.

If the speech signal is preemphasized prior to LP analysis to compensate for the effects of radiation and the nonwhite glottal pulse, then the resulting cross-sectional areas are often similar to the human vocal tract configuration used to produce the speech under analysis (40). They cannot be guaranteed to match, however, because of the nonuniqueness properties of the vocal-tract configuration. For example, to keep their lip opening small, ventriloquists exploit this property by compensating with the remainder of their vocal tract configuration.

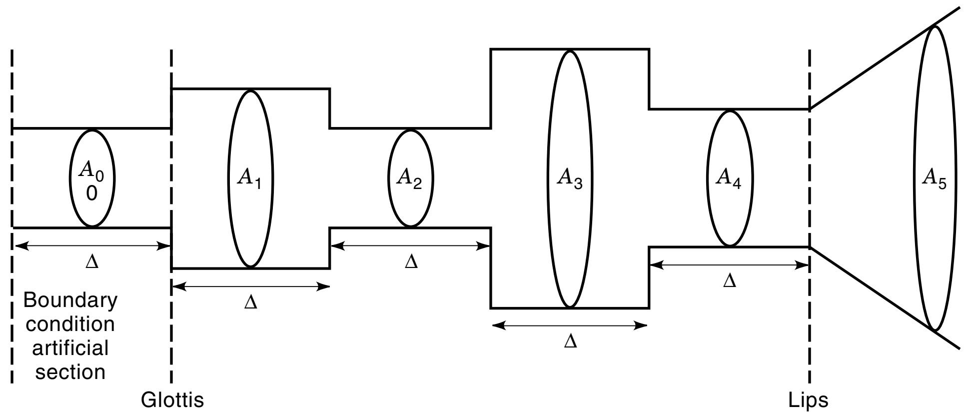

Figure 5. Acoustic tube model of speech production.

Narrow bandwidth poles result in $|k_i| \sim 1$. 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

$$g_i = \log\left[\frac{A_{i+1}}{A_i}\right] = \log\left[\frac{1+k_i}{1-k_i}\right] = 2\tanh^{-1} k_i \vee i = 1, 2, \ldots, p \tag{15}$$

Arcsin Reflection Coefficients. To avoid the singularity of the LARs at $k_i = 1$ while retaining approximately uniform spectral sensitivity, the arcsin of the RCs are a common choice

$$g'_i = \sin^{-1} k_i \vee i = 1, 2, \ldots, p \tag{16}$$

Line Spectrum Pair Frequencies. The LSPs are a representation of the PCs 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) (39)

$$\begin{aligned} A(z) &= \tfrac{1}{2}[P(z) + Q(z)] \\ P(z) &= A(z) + z^{-(p+1)}A(z^{-1}) \\ Q(z) &= A(z) - z^{-(p+1)}A(z^{-1}) \end{aligned} \tag{17}$$

where the LSPs are the frequencies of the zeros of $P(z)$ and $Q(z)$. By definition, a stable LP synthesis filter has all its poles inside the unit circle in the z plane. The corresponding inverse filter is therefore minimum phase inverse because it has no poles or zeros outside the unit circle. Any minimum phase polynomial can be mapped by this transform to represent each of its roots by a pair of frequencies (phases) with unit magnitude. The LSP representation of the LP filter has a direct frequency-domain interpretation that is especially useful in efficient (accurate and compact) coding and smoothing of the LP filter coefficients (41).

For example, an eighth-order 8 kHz LP analysis of the vowel /u/ (as in foot) had the predictor coefficients shown in Table 4. 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 (Fig. 6). These resonance frequencies are in agreement with the Peterson and Barney formant frequency data for the vowel /u/ (40).

Because the PCs are real, the Fundamental Theorem of Algebra guarantees that the roots of $A(z)$, $P(z)$, and $Q(z)$ will occur in complex conjugate pairs. Because of this conjugate property, the bottom half of the z plane is redundant. The LSPs at zero and π are always present by construction of P and Q. Therefore, the PCs can be represented by the number of LSPs equal to the prediction order p and are represented by the frequencies of the zeros of P and Q in the top-half z plane (Fig. 7).

The LSPs satisfy an interlacing property of the zeros of the P and Q polynomials, which holds for all minimum phase $A(z)$ polynomials (39)

$$0 = \omega_0^{(Q)} < \omega_1^{(P)} < \omega_2^{(Q)} < \cdots < \omega_{p-1}^{(P)} < \omega_p^{(Q)} < \omega_{p+1}^{(P)} = \pi \tag{18}$$

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 between the corresponding LSPs. 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. Figures 6 and 7 demonstrate this behavior.

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. The mel-warping transforms the frequency scale

Table 4. Example of Eighth-Order Linear Predictor Coefficients for the Vowel /u/ as in "Foot"

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

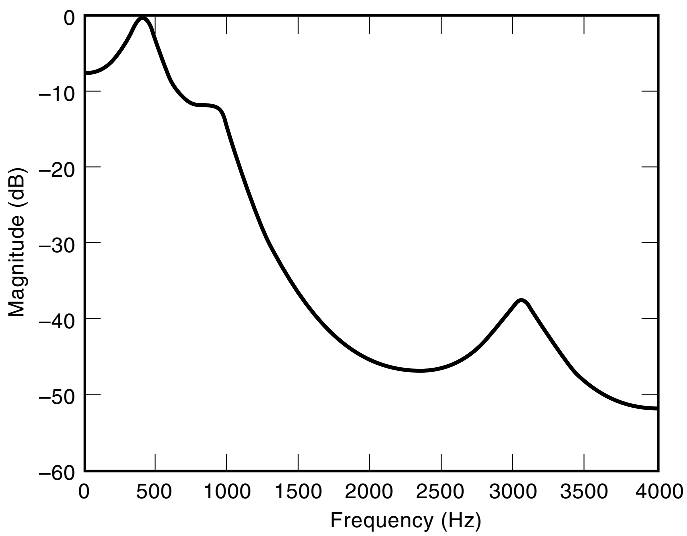

Figure 6. Frequency response for the vowel /u/.

to place less emphasis on high frequencies. It is based on the nonlinear human perception of the frequency of sounds (42). 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 (9). Perhaps the most compelling reason for using the mel-warped cepstrums is that it has been demonstrated to work well in speaker recognition systems (28) and, somewhat ironically, in speech recognition systems (42), too.

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.

FEATURE SELECTION AND MEASURES

To apply mathematical tools without loss of generality, the speech signal can be represented by a sequence of feature vectors. In this section, the selection of appropriate features is discussed, along with methods to estimate (extract or measure) them. This is known as feature selection and feature extraction.

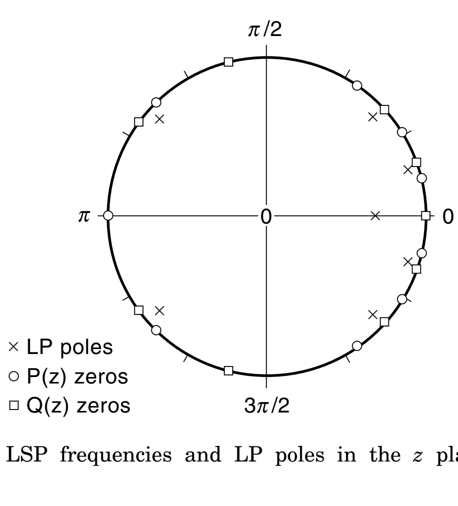

Figure 7. LSP frequencies and LP poles in the z plane for the vowel/u/.

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 the state of the art in discriminant analysis is given in Gnanadesikan and Kettenring (43).

Traditional Feature Selection

Feature extraction is the estimation of variables, called a feature vector, from another set of variables (e.g., an observed speech signal time series). Feature selection is the transformation of these observation vectors to feature vectors. The goal of feature selection is to find a transformation to a relatively low-dimensional feature space that preserves the information pertinent to the application, while enabling meaningful comparisons to be performed using simple measures of similarity.

Although it might be tempting at first to select all the extracted features, the "curse of dimensionality" quickly becomes overwhelming (44). As more features are used, the feature dimensions increase, which imposes severe requirements on computation and storage in both training and testing. The demand for a large amount of training data to represent a speaker's voice characteristics grows exponentially with the dimension of the feature space. This severely restricts the usefulness of nonparametric procedures (no assumed underlying statistical model) and higher-order transforms.

The traditional statistical methods to reduce dimensionality, and avoid this curse, are principal component analysis and factor analysis. Principal component analysis seeks to find a lower-dimensional representation that accounts for variance of the features. Factor analysis seeks to find a lower-dimensional representation that accounts for correlations among the features. In other disciplines, principal component analysis is called the *Karhunen–Loève expansion* (KLE) or *eigenvector orthonormal expansion.* Because each eigenvector can be ranked by its corresponding eigenvalue, a subset of the eigenvectors can be chosen to minimize the MSE in representing the data. Although KLE is optimum for representing classes with the same mean, it is not necessarily optimum for discriminating between classes (45). Because speaker recognition is a discrimination problem, as opposed to a representation problem, we seek other means to reduce the dimensionality of the data.

Linear transformation are capable of dividing the feature space by a hyperplane. If data are linearly separable, then it can be discriminated by a hyperplane. In the case of a two-dimensional feature space, the hyperplane collapses to a line. As shown in Eq. (19), given a random variable $\mathbf{x}$ distributed normally with mean $\boldsymbol{\mu}_x$ and covariance $\mathbf{C}_x$ and an $m \times n$

transformation matrix $\mathbf{A}$, $p(\mathbf{x}) \sim N(\boldsymbol{\mu}_x, \mathbf{C}_x)$, $\mathbf{y} = \mathbf{A}\mathbf{x}$ is an m-component feature vector and $p(\mathbf{y}) \sim N(\mathbf{A}\boldsymbol{\mu}_x, \mathbf{A}\mathbf{C}_x\mathbf{A}^T)$, where T denotes matrix transpose

$$\begin{aligned}
\mathbf{y} &= \mathbf{A}\mathbf{x} \\
\boldsymbol{\mu}_y &= E[\mathbf{y}] = E[\mathbf{A}\mathbf{x}] = \mathbf{A}E[\mathbf{x}] \\
&= \mathbf{A}\boldsymbol{\mu}_x \\
\mathbf{C}_y &= E[(\mathbf{y}-\boldsymbol{\mu}_y)(\mathbf{y}-\boldsymbol{\mu}_y)^T] = E[\mathbf{A}(\mathbf{x}-\boldsymbol{\mu}_x)(\mathbf{x}-\boldsymbol{\mu}_x))^T] \\
&= E[\mathbf{A}(\mathbf{x}-\boldsymbol{\mu}_x)(\mathbf{x}-\boldsymbol{\mu}_x)^T\mathbf{A}^T] = \mathbf{A}E[(\mathbf{x}-\boldsymbol{\mu}_x)(\mathbf{x}-\boldsymbol{\mu}_x)^T]\mathbf{A}^T \\
&= \mathbf{A}\mathbf{C}_x\mathbf{A}^T
\end{aligned} \tag{19}$$

Thus, a linear transformation of a multivariate normal vector also has a normal density. Any linear combination of normally distributed random variables is again normal. This can be used to tremendous advantage if the feature densities of the speakers are assumed to be normal. This allows us to lump all the other speaker probability density functions (pdfs) into a single, normal pdf. Thus, pairwise (two-class) discriminators can be designed to separate the claimant speaker from other speakers.

In the special case where the transformation is a unit length vector $\mathbf{a}$, $y = \mathbf{a}\mathbf{x}$ is a scalar that represents the projection of $\mathbf{x}$ onto a line in the direction of $\mathbf{a}$. In general, $\mathbf{A}\mathbf{C}_x\mathbf{A}^T$ is the variance of the projection of $\mathbf{x}$ onto the column space of $\mathbf{A}$. Thus, knowledge of the covariance matrix allows us to calculate the dispersion of the data in any direction.

In Fig. 8, two classes are represented by boxes and circles in a two-dimensional feature space (x_1, x_2). Here, we see that if feature x_1 or x_2 were used by itself, discrimination errors would occur because of the overlap between the projected classes onto the x_1 or x_2 axes. It is quite clear, however, that the data are perfectly linearly separable by the dashed line. If the data are linearly transformed onto the column space of $\mathbf{A}$, perfect discrimination is achieved. In addition, we can see a clustering effect by the reduced variance of the projection onto the column space of $\mathbf{A}$.

Note that data may not always be discriminated well by a linear transformation. In these cases, a nonlinear transformation may lead to improved discrimination. An example is the classes defined by the members of interlocking spirals. No line can separate the spirals, but a nonlinear transformation could yield perfect discrimination.

The goal of speaker-recognition feature selection is to find a set that minimizes the probability of error. Unfortunately, an explicit mathematical expression is unavailable, except for trivial cases, which hinders rigorous mathematical development. Even for normal pdfs, a numerical integration is required to determine probability of error (except for the equal covariance case) (46).

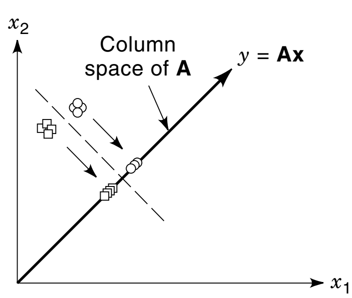

Figure 8. Linear transformation with perfect discrimination.

To make the problem mathematically tractable, one approach is to select a feature set that exhibits low intraspeaker variability and high interspeaker variability. A technique that can be used to find good features is analysis of variance (ANOVA), which involves measuring Fisher's F-ratio, Eq. (20), between the sample pdfs of different features. For speaker verification, high F-ratios are desirable.

$$F = \frac{\text{Variance of speaker means}}{\text{Average intraspeaker variance}} \tag{20}$$

Unfortunately, ANOVA requires evaluating the F-ratio for many different *combinations* of features to be really useful. For example, two features with high individual F-ratios might be highly correlated and as a feature vector, less effective than two features that individually have low F-ratios. The usefulness of the F-ratio as a discrimination measure is further reduced if the classes are multimodal or if they have the same means. This is a fatal flaw with any criterion that is dominated by differences between class means. This will now be demonstrated.

Normal Density with Equal Means. The normal pdf is often a good approximation to real-world density functions. Classes will exhibit normal densities when each pattern of a class is a random vector formed by superposition of a random vector upon a nonrandom vector, where the superimposed random vectors are drawn from the same normal density. This is a good approximation to real-world situations characterized by independent identically distributed additive Gaussian noise. The normal pdf has some striking advantages. It is one of the simplest parametric models, being characterized by a mean and variance. In addition, the sum of normal random variables yields a normal random variable.

The *n-variate normal pdf* is defined as

$$\begin{aligned}
p(\mathbf{x}) &= (2\pi)^{-n/2}|\mathbf{C}|^{-1/2}\exp[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})] \\
&\sim N(\boldsymbol{\mu}, \mathbf{C})
\end{aligned} \tag{21}$$

where $\mathbf{C}$ is the $n \times n$ covariance matrix, and $\boldsymbol{\mu}$ is an n-dimensional column component mean vector. Note that in Eq. (21), contours of constant probability occur for values of $\mathbf{x}$ where the argument of the exponential is constant. Neglecting the scaling factor of $-(1/2)$, the argument of the exponential is referred to as the *Mahalanobis distance* d_{M}^2 between $\mathbf{x}$ and $\boldsymbol{\mu}$

$$d_{\mathrm{M}}^2 = (\mathbf{x}-\boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \tag{22}$$

Thus, the loci of points of constant density are hyperellipsoids of constant Mahalanobis distance to $\boldsymbol{\mu}$. The principal axes of these hyperellipsoids are given by the eigenvectors of $\mathbf{C}$, and their eigenvalues determine the lengths of the corresponding axes.

Samples drawn from a multivariate normal density tend to cluster. The center of the cluster is determined by the mean and the shape of the cluster is determined by the covariance matrix. In the bivariate (n = 2) case, it is convenient for visualization to show the 1-sigma ellipse. The 1-sigma ellipse is centered on the means, its major axes are determined by the 1-sigma standard deviations, and its orientation is deter-

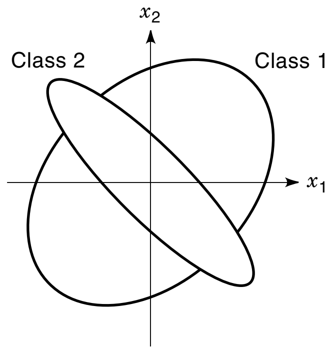

Figure 9. Unequal covariance.

mined by the covariance between the variables. For example, Fig. 9 shows the bivariate 1-sigma ellipses for two classes with equal means, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = [0\ 0]^T$ and unequal covariance matrices.

Although there is no line that can perfectly discriminate these two classes, it's easy to visualize that a 45° projection would provide some discrimination power. However, the F-ratio would indicate that these features, x_1 and x_2, are powerless because the classes have the same means in the $x_1 - x_2$ space.

Now consider a bimodal pdf. Figure 10 shows class 1 as being bimodal in x_1. The means of both classes are the same; hence, the F-ratio would show feature x_1 as powerless. It is clear from Fig. 10, however, that x_1 is powerful because significant discriminatory information exists along feature x_1. Thus, caution should be used with any criterion, such as the F-ratio, that relies on class means. If the classes have the same means or are not unimodal, the F-ratio can be a poor measure of discrimination power. Clearly, we seek a criterion that more accurately portrays discrimination power.

Mean and Covariance Estimation

The unbiased estimate (UBE) of the covariance is given by the sample covariance

$$\hat{\mathbf{C}} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \tag{23}$$

The UBE and maximum likelihood estimate (MLE) of covariance differ only by their scaling factors of $1/(N-1)$ and $1/N$, respectively, and they are both referred to as sample covariance matrices. When the mean is being estimated too, the UBE is generally preferred; however, they are practically identical when N is large.

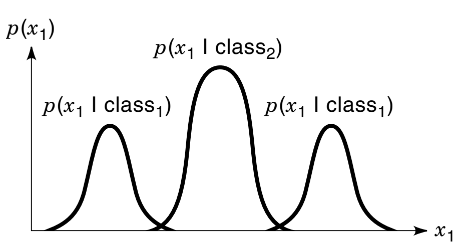

Figure 10. A bimodal class.

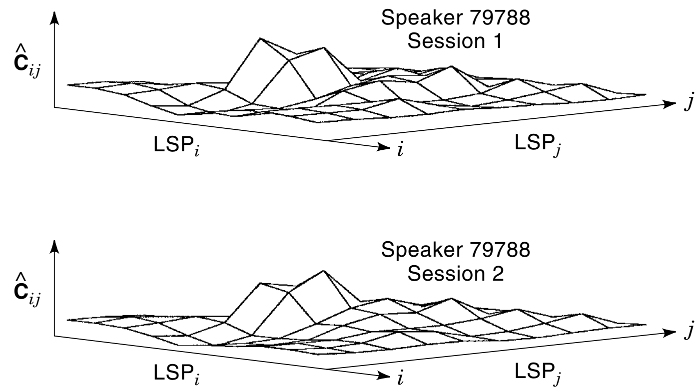

Figure 11. LSP covariance matrices: Different sessions, same speaker.

To estimate the mean and covariance when all samples are not yet available or when dealing with a large number of samples, recursive computation methods are desirable. Denoting an estimate based upon N samples as $\hat{\boldsymbol{\mu}}_N$ and on $N + 1$ samples as $\hat{\boldsymbol{\mu}}_{N+1}$, the sample mean is

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{N+1} &= \frac{1}{N+1}\sum_{k=1}^{N+1}\mathbf{x}_k \\ &= \hat{\boldsymbol{\mu}}_N + \frac{1}{N+1}(\mathbf{x}_{N+1} - \hat{\boldsymbol{\mu}}_N)\end{aligned} \tag{24}$$

Similarly, the UBE sample covariance matrix recursion $\hat{\mathbf{C}}_{N+1}$ is

$$\begin{aligned}\hat{\mathbf{C}}_{N+1} &= \frac{1}{N}\sum_{k=1}^{N+1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{N+1})(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{N+1})^T \\ &= \frac{N-1}{N}\hat{\mathbf{C}}_N + \frac{1}{N+1}(\mathbf{x}_{N+1} - \hat{\boldsymbol{\mu}}_N)(\mathbf{x}_{N+1} - \hat{\boldsymbol{\mu}}_N)^T\end{aligned} \tag{25}$$

Sample covariance matrices using LSP features are shown in the mesh plots of Figs. 11 and 12. In each plot, the variances and covariances of 10 LSP coefficients are represented in the vertical direction on a 10 × 10 mesh. From a total of 80 s of speech, each matrix (mesh plot) was generated from the LSP vectors corresponding to voiced speech. Notice that

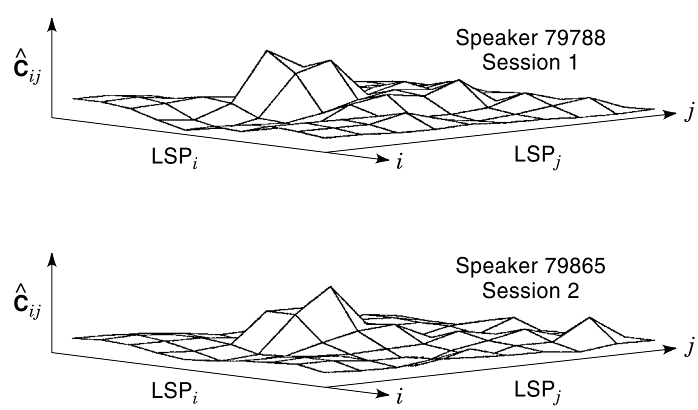

Figure 12. LSP covariance matrices: Different speakers.

these covariance matrices for different sessions of the same speaker appear to be similar.

These LSP covariance matrices appear to have more differences between speakers than similarities for the same speaker. As shown later, the LSP covariance matrices can capture speaker identity.

Divergence Measure

Divergence is a measure of dissimilarity between two classes based upon information theory (47). It provides a means of feature ranking and evaluation of class discrimination effectiveness. The following development is based upon Tou and Gonzalez's derivation (45). Let the *likelihood* of occurrence of pattern $\mathbf{x}$, given that it belongs to class ω_i, be

$$p_i(\mathbf{x}) = p(\mathbf{x}|\omega_i) \tag{26}$$

and likewise for class ω_j,

$$p_j(\mathbf{x}) = p(\mathbf{x}|\omega_j) \tag{27}$$

Then, the *discriminating information* of an observation $\mathbf{x}$, in the Bayes classifier sense, for class ω_i versus class ω_j can be measured by the logarithm of the *likelihood ratio*

$$u_{ij} = \ln \frac{p_i(\mathbf{x})}{p_j(\mathbf{x})} \tag{28}$$

Entropy is the statistical measure of information or uncertainty. The *population entropy* H for a given ensemble of pattern vectors having a pdf $p(\mathbf{x})$ is the expectation

$$\begin{aligned} H &= -E[\ln p(\mathbf{x})] \\ &= -\int_x p(\mathbf{x}) \ln p(\mathbf{x})\, dx \end{aligned} \tag{29}$$

Similarly, the entropy of the ith class of population of patterns is

$$H_1 = -\int_x p_i(\mathbf{x}) \ln p_i(\mathbf{x})\, dx \tag{30}$$

The *average discriminating information* for class ω_i versus class ω_j over all observations, also known as *directed divergence, Kullback–Leibler number* (47) or *discrimination* (48), is then

$$\begin{aligned} I(i,j) &= \int_x p_i(\mathbf{x}) u_{ij}\, dx \\ &= \int_x p_i(\mathbf{x}) \ln \frac{p_i(\mathbf{x})}{p_j(\mathbf{x})}\, dx \end{aligned} \tag{31}$$

Likewise, the discriminating information for class ω_j versus class ω_i can be measured by the logarithm of the likelihood ratio

$$u_{ji} = \ln \frac{p_j(\mathbf{x})}{p_i(\mathbf{x})} \tag{32}$$

The average discriminating information for class ω_j is then

$$I(j,i) = \int_x p_j(\mathbf{x}) \ln \frac{p_j(\mathbf{x})}{p_i(\mathbf{x})}\, dx \tag{33}$$

The *divergence* (the symmetric directed divergence) is defined as the total average information for discriminating class ω_i from class ω_j

$$\begin{aligned} J_{ij} &= I(i,j) + I(j,i) \\ &= \int_x [p_i(\mathbf{x}) - p_j(\mathbf{x})] \ln \frac{p_i(\mathbf{x})}{p_j(\mathbf{x})}\, dx \end{aligned} \tag{34}$$

Now, to select features with this measure, we need the feature pdf for each pattern class. Assuming the pattern classes are n-variate normal populations

$$p_i(\mathbf{x}) \sim N(\boldsymbol{\mu}_i, \mathbf{C}_i) \qquad p_j(\mathbf{x} \sim N(\boldsymbol{\mu}_j, \mathbf{C}_j) \tag{35}$$

Substituting Eq. (21) into Eq. (28) yields the log likelihood ratio

$$\begin{aligned} u_{ij} &= \tfrac{1}{2} \ln \frac{|\mathbf{C}_j|}{|\mathbf{C}_i|} - \tfrac{1}{2}\mathrm{tr}[\mathbf{C}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T] \\ &\quad + \tfrac{1}{2}\mathrm{tr}[\mathbf{C}_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)(\mathbf{x} - \boldsymbol{\mu}_j)^T] \end{aligned} \tag{36}$$

where tr is the matrix trace function. The average information for discrimination between these two classes is

$$\begin{aligned} I(i,j) &= \int_x p_i(\mathbf{x}) u_{ij}\, dx \\ &= \int_x (2\pi)^{-n/2} |\mathbf{C}_i|^{-1/2} \exp[-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \mathbf{C}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)] \\ &\quad \times \left\{ \tfrac{1}{2} \ln \frac{|\mathbf{C}_j|}{|\mathbf{C}_i|} - \tfrac{1}{2}\mathrm{tr}[\mathbf{C}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T] \right. \\ &\quad \left. + \tfrac{1}{2}\mathrm{tr}[\mathbf{C}_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)(\mathbf{x} - \boldsymbol{\mu}_j)^T] \right\} dx \\ &= \tfrac{1}{2} \ln \frac{|\mathbf{C}_j|}{|\mathbf{C}_i|} + \tfrac{1}{2}\mathrm{tr}[\mathbf{C}_i(\mathbf{C}_j^{-1} - \mathbf{C}_i^{-1})] \\ &\quad + \tfrac{1}{2}\mathrm{tr}[\mathbf{C}_j^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T] \end{aligned} \tag{37}$$

Let the difference in the means be represented as

$$\boldsymbol{\delta} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \tag{38}$$

The average information for discrimination between these two classes is

$$I(i,j) = \tfrac{1}{2} \ln \frac{|\mathbf{C}_j|}{|\mathbf{C}_i|} + \tfrac{1}{2}\mathrm{tr}[\mathbf{C}_i(\mathbf{C}_j^{-1} - \mathbf{C}_i^{-1})] + \tfrac{1}{2}\mathrm{tr}[\mathbf{C}_j^{-1}\boldsymbol{\delta}\boldsymbol{\delta}^T] \tag{39}$$

Hence, the *divergence* for these two normally distributed classes is

$$\begin{aligned} J_{ij} &= \tfrac{1}{2} \ln \frac{|\mathbf{C}_j|}{|\mathbf{C}_i|} + \tfrac{1}{2}\mathrm{tr}[\mathbf{C}_i(\mathbf{C}_j^{-1} - \mathbf{C}_i^{-1})] \\ &\quad + \tfrac{1}{2}\mathrm{tr}[\mathbf{C}_j^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T] \\ &\quad + \tfrac{1}{2} \ln \frac{|\mathbf{C}_i|}{|\mathbf{C}_j|} + \tfrac{1}{2}\mathrm{tr}[\mathbf{C}_j(\mathbf{C}_i^{-1} - \mathbf{C}_j^{-1})] \\ &\quad + \tfrac{1}{2}\mathrm{tr}[\mathbf{C}_i^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)^T] \\ &= \tfrac{1}{2}\mathrm{tr}[(\mathbf{C}_i - \mathbf{C}_j)(\mathbf{C}_j^{-1} - \mathbf{C}_i^{-1})] \\ &\quad + \tfrac{1}{2}\mathrm{tr}[(\mathbf{C}_i^{-1} + \mathbf{C}_j^{-1})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T] \\ &= \tfrac{1}{2}\mathrm{tr}[(\mathbf{C}_i - \mathbf{C}_j)(\mathbf{C}_j^{-1} - \mathbf{C}_i^{-1})] + \tfrac{1}{2}\mathrm{tr}[(\mathbf{C}_i^{-1} + \mathbf{C}_j^{-1})\boldsymbol{\delta}\boldsymbol{\delta}^T] \end{aligned} \tag{40}$$

Divergence Shape. Note that Eq. (40) is the sum of two components, one based solely upon differences between the covariance matrices and the other involves differences between the mean vectors, $\boldsymbol{\delta}$. These components can be characterized, respectively, as differences in shape and size of the pdfs. This shape component, the *divergence shape,* will prove very useful later on

$$J'_{ij} = \mathrm{tr}[(\mathbf{C}_i - \mathbf{C}_j)(\mathbf{C}_j^{-1} - \mathbf{C}_i^{-1})] \tag{41}$$

Equation (40) is slightly complicated, so let us consider two simplifying special cases.

Equal Covariance Divergence. First, for the equal covariance case, let

$$\mathbf{C}_i = \mathbf{C}_j = \mathbf{C} \tag{42}$$

This leaves only the last term from Eq. (37)

$$\begin{aligned} I(i,j) &= \tfrac{1}{2}\mathrm{tr}[\mathbf{C}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T] \\ &= \tfrac{1}{2}\mathrm{tr}[\mathbf{C}^{-1}\boldsymbol{\delta}\boldsymbol{\delta}^T] \\ &= \tfrac{1}{2}\boldsymbol{\delta}^T\mathbf{C}^{-1}\boldsymbol{\delta} \end{aligned} \tag{43}$$

and, therefore,

$$\begin{aligned} J_{ij} &= \tfrac{1}{2}\mathrm{tr}[\mathbf{C}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T] \\ &\quad + \tfrac{1}{2}\mathrm{tr}[\mathbf{C}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)^T] \\ &= \mathrm{tr}[\mathbf{C}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T] \\ &= \boldsymbol{\delta}^T\mathbf{C}^{-1}\boldsymbol{\delta} \end{aligned} \tag{44}$$

Comparing this with Eq. (22), the divergence for this normal equal covariance case is simply the Mahalanobis distance between the two class means.

For a univariate ($n = 1$) normal equal variance σ^2 population

$$I(i,j) = \frac{1}{2}\frac{(\mu_i - \mu_j)^2}{\sigma^2} \tag{45}$$

Reassuringly, the divergence in this equal covariance case is the familiar F-ratio

$$J_{ij} = \frac{(\mu_i - \mu_j)^2}{\sigma^2} \tag{46}$$

Equal Mean Divergence. Next, for the equal population means case,

$$\boldsymbol{\mu}_i = \boldsymbol{\mu}_j \quad \boldsymbol{\delta} = 0 \tag{47}$$

The average information is

$$\begin{aligned} I(i,j) &= \tfrac{1}{2}\ln\frac{|\mathbf{C}_j|}{|\mathbf{C}_i|} + \tfrac{1}{2}\mathrm{tr}[\mathbf{C}_i(\mathbf{C}_j^{-1} - \mathbf{C}_i^{-1})] \\ &= \tfrac{1}{2}\ln\frac{|\mathbf{C}_j|}{|\mathbf{C}_i|} + \tfrac{1}{2}\mathrm{tr}[\mathbf{C}_i\mathbf{C}_j^{-1}] - \frac{n}{2} \end{aligned} \tag{48}$$

The divergence is

$$\begin{aligned} J_{ij} &= \tfrac{1}{2}\mathrm{tr}[(\mathbf{C}_i - \mathbf{C}_j)(\mathbf{C}_j^{-1} - \mathbf{C}_i^{-1})] \\ &= \tfrac{1}{2}\mathrm{tr}[\mathbf{C}_i\mathbf{C}_j^{-1}] + \mathrm{tr}[\mathbf{C}_j\mathbf{C}_j^{-1}] - n \end{aligned} \tag{49}$$

Divergence Properties. The divergence satisfies all the metric properties except the triangle inequality. Thus, divergence is not termed a distance (49). The following properties of divergence are proven in the landmark paper of Kullback and Leibler (49). Positivity (i.e., almost positive definite) and symmetry properties are satisfied

$$\begin{aligned} &J_{ij} \geq 0 \quad \text{and} \quad J_{ij} = 0 \,\text{iff}\, p_i \neq p_j \\ &J_{ij} = J_{ji} \end{aligned} \tag{50}$$

By counterexample, divergence can be shown to violate the triangle inequality by taking $p_1 \sim N(0, 1)$, $p_2 \sim N(0, 4)$, and $p_3 \sim N(0, 5)$; thus, $J_{13} > J_{12} + J_{23}$.

Additional measurements (increased dimensionality) cannot decrease divergence

$$J_{ij}(x_1, x_2, K, x_m) \leq J_{ij}(x_1, x_2, \ldots, x_m, x_{m+1}) \tag{51}$$

As should be expected from an information-theoretic measure, processing cannot increase divergence (48). Thus, transformation of the feature space must maintain or decrease divergence. Furthermore, divergence can be shown to be invariant under *onto* measurable transformation (49). Kullback's real-analysis-based proof is rather difficult to follow, so let us consider the special case of proving the invariance of the divergence measure under nonsingular linear transformation (affine transformation could be similarly shown)

$$\begin{aligned} &\text{if}\; p(\mathbf{x}) \sim N(\boldsymbol{\mu}_x, \mathbf{C}_x) \;\text{where}\; \mathbf{x} \in \Re^n \quad \text{and} \quad \mathbf{A} \in \Re^{m \times n} \\ &\text{let}\; \mathbf{y} = \mathbf{A}\mathbf{x} \quad \text{where} \quad \mathbf{y} \in \Re^m \\ &\text{then}\; \boldsymbol{\mu}_y = E[y] = E[\mathbf{A}\mathbf{x}] = \mathbf{A}E[\mathbf{x}] = \mathbf{A}\boldsymbol{\mu}_x \\ &\mathbf{C}_y = E[(y - \boldsymbol{\mu}_y)(y - \boldsymbol{\mu}_y)^T] \\ &\quad = E[(\mathbf{A}\mathbf{x} - \mathbf{A}\boldsymbol{\mu}_x)(\mathbf{A}\mathbf{x} - \mathbf{A}\boldsymbol{\mu}_x)^T] = \mathbf{A}\mathbf{C}_x\mathbf{A}^T \\ &\therefore p(y) \sim N(\mathbf{A}\boldsymbol{\mu}_x, \mathbf{A}\mathbf{C}_x\mathbf{A}^T) \\ &\text{let}\; J_{ij}^{(x)} = \tfrac{1}{2}\mathrm{tr}[(\mathbf{C}_i^{(x)} - \mathbf{C}_j^{(x)})((\mathbf{C}_j^{(x)})^{-1} - (\mathbf{C}_i^{(x)})^{-1})] \\ &\quad + \tfrac{1}{2}\mathrm{tr}[((\mathbf{C}_i^{(x)})^{-1} + (\mathbf{C}_j^{(x)})^{-1})(\mu_i^{(x)} - \boldsymbol{\mu}_j^{(x)})(\boldsymbol{\mu}_i^{(x)} - \boldsymbol{\mu}_j^{(x)})^T] \\ &\text{then}\; J_{ij}^{(y)} = \tfrac{1}{2}\mathrm{tr}[(\mathbf{A}\mathbf{C}_i^{(x)}\mathbf{A}^T - \mathbf{A}\mathbf{C}_j^{(x)}\mathbf{A}^T) \\ &\quad \cdot ((\mathbf{A}^T)^{-1}(\mathbf{C}_j^{(x)})^{-1}\mathbf{A}^{-1} - (\mathbf{A}^T)^{-1}(\mathbf{C}_i^{(x)})^{-1}\mathbf{A}^{-1})] \\ &\quad + \tfrac{1}{2}\mathrm{tr}[((\mathbf{A}^T)^{-1}(\mathbf{C}_i^{(x)})^{-1}\mathbf{A}^{-1} + (\mathbf{A}^T)^{-1}(\mathbf{C}_j^{(x)})^{-1}\mathbf{A}^{-1}) \\ &\quad \cdot (\mathbf{A}\boldsymbol{\mu}_i^{(x)} - \mathbf{A}\boldsymbol{\mu}_j^{(x)})(\mathbf{A}\boldsymbol{\mu}_i^{(x)} - \mathbf{A}\boldsymbol{\mu}_j^{(x)})^T] \\ &= \tfrac{1}{2}\mathrm{tr}[\mathbf{A}(\mathbf{C}_i^{(x)} - \mathbf{C}_j^{(x)})\mathbf{A}^T(\mathbf{A}^T)^{-1}((\mathbf{C}_j^{(x)})^{-1} \\ &\quad - (\mathbf{C}_j^{(x)})^{-1})\mathbf{A}^{-1}] \\ &\quad + \tfrac{1}{2}\mathrm{tr}[(\mathbf{A}^T)^{-1}((\mathbf{C}_i^{(x)})^{-1} + (\mathbf{C}_j^{(x)})^{-1})\mathbf{A}^{-1}\mathbf{A}(\boldsymbol{\mu}_i^{(x)} - \boldsymbol{\mu}_j^{(x)}) \\ &\quad \cdot (\mathbf{A}(\boldsymbol{\mu}_i^{(x)} - \boldsymbol{\mu}_j^{(x)}))^T] \\ &= \tfrac{1}{2}\mathrm{tr}[\mathbf{A}\mathbf{A}^{-1}(\mathbf{C}_i^{(x)} - \mathbf{C}_j^{(x)})((\mathbf{C}_j^{(x)})^{-1} - (\mathbf{C}_i^{(x)})^{-1})] \\ &\quad + \tfrac{1}{2}\mathrm{tr}[(\mathbf{A}^T)^{-1}\mathbf{A}^T((\mathbf{C}_i^{(x)})^{-1} \\ &\quad + (\mathbf{C}_j^{(x)})^{-1})(\boldsymbol{\mu}_i^{(x)} - \boldsymbol{\mu}_j^{(x)})(\boldsymbol{\mu}_j^{(x)})^T] \\ &= J_{ij}^{(x)} \end{aligned} \tag{52}$$

This is a powerful result because of the many useful linear transformations (e.g., discrete Fourier transform, discrete cosine transform, and discrete convolution). For example, if the frequency domain can be attained via linear transformation, there is no need to separately consider this mapping of the features. This invariance also implies that linear feature selection is unnecessary unless dimensionality reduction is desired.

Divergence is additive for independent measurements

$$J_{ij}(x_1, x_2, \ldots, x_m) = \sum_{k=1}^{m} J_{ij}(x_k) \tag{53}$$

This allows ranking the importance of each feature according to its associated divergence.

Example of Equal Covariance Divergence. The preceding concepts are demonstrated here based upon an example taken from Tou and Gonzalez (45). Intermediate steps have been added to aid the reader. Given the observations of Eq. (54)

$$\begin{aligned}
\mathbf{x}_{11} &= \begin{bmatrix}0\\0\\0\end{bmatrix} \mathbf{x}_{12} = \begin{bmatrix}1\\0\\0\end{bmatrix} \mathbf{x}_{13} = \begin{bmatrix}1\\0\\1\end{bmatrix} \mathbf{x}_{14} = \begin{bmatrix}1\\1\\0\end{bmatrix}\\
\mathbf{x}_{21} &= \begin{bmatrix}0\\0\\1\end{bmatrix} \mathbf{x}_{22} = \begin{bmatrix}0\\1\\0\end{bmatrix} \mathbf{x}_{23} = \begin{bmatrix}0\\1\\1\end{bmatrix} \mathbf{x}_{24} = \begin{bmatrix}1\\1\\1\end{bmatrix}
\end{aligned} \tag{54}$$

where the first index indicates class ω_1 or ω_2. These patterns are shown in Fig. 13. From this figure, it is obvious that the data could be perfectly discriminated by a plane slicing through the data. Let us see how the divergence measure separates the classes.

To estimate the population means, we approximate the mean vectors by the sample average over N samples

$$\begin{aligned}
\boldsymbol{\mu} &= E[\mathbf{x}]\\
&= \int_x \mathbf{x}p(\mathbf{x})\,dx\\
&\approx \frac{1}{N}\sum_{j=1}^{N} \mathbf{x}_j
\end{aligned} \tag{55}$$

If the mean is not considered a random variable, the covariance may be similarly estimated using a sample average

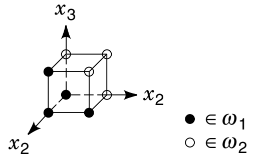

Figure 13. Original observation vectors [after Tou and Gonzalez (45)].

$$\begin{aligned}
\mathbf{C} &= E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T]\\
&= E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}^T-\boldsymbol{\mu}^T)]\\
&= E[\mathbf{x}\mathbf{x}^T - \mathbf{x}\boldsymbol{\mu}^T - \boldsymbol{\mu}\mathbf{x}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T]\\
&= E[\mathbf{x}\mathbf{x}^T - 2\mathbf{x}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T]\\
&= E[\mathbf{x}\mathbf{x}^T] - 2E[\mathbf{x}\boldsymbol{\mu}^T] + E[\boldsymbol{\mu}\boldsymbol{\mu}^T]\\
&= E[\mathbf{x}\mathbf{x}^T] - 2\boldsymbol{\mu}\boldsymbol{\mu}T + \boldsymbol{\mu}\boldsymbol{\mu}^T\\
&= E[\mathbf{x}\mathbf{x}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T\\
&\approx -\boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j\mathbf{x}_j^T
\end{aligned} \tag{56}$$

For each class, plugging in the observation vectors, we find that the means are unequal and the covariances are equal

$$\boldsymbol{\mu}_1 = \frac{1}{4}\begin{bmatrix}3\\1\\1\end{bmatrix} \quad \boldsymbol{\mu}_2 = \frac{1}{4}\begin{bmatrix}1\\3\\3\end{bmatrix} \quad \mathbf{C} = \mathbf{C}_1 = \mathbf{C}_2 = \frac{1}{16}\begin{bmatrix}3 & 1 & 1\\1 & 3 & -1\\1 & -1 & 3\end{bmatrix} \tag{57}$$

$$\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \frac{1}{4}\begin{bmatrix}2\\-2\\-2\end{bmatrix} \quad \mathbf{C}^{-1} = \begin{bmatrix}8 & -4 & -4\\-4 & 8 & 4\\-4 & 4 & 8\end{bmatrix} \tag{58}$$

To maximize divergence in this special case, choose the transformation matrix as the transpose of the nonzero eigenvalue's corresponding eigenvector of $\mathbf{C}^{-1}\delta\delta^{\mathrm{T}}$ (a closed-form solution does not exist for the general case) (50)

$$\mathbf{C}^{-1}\boldsymbol{\delta}\boldsymbol{\delta}^T = \frac{1}{4}\begin{bmatrix}1 & -1 & -1\\-1 & 1 & 1\\-1 & 1 & 1\end{bmatrix} \tag{59}$$

$$\lambda = \frac{3}{4} \quad \mathbf{e} = \begin{bmatrix}-1\\1\\1\end{bmatrix} \tag{60}$$

$$\mathbf{A} = \mathbf{e}^T = [-1 \quad 1 \quad 1] \tag{61}$$

$$y = \mathbf{A}\mathbf{x} \tag{62}$$

$$\begin{aligned}
&y_{11} = 0 \quad y_{12} = -1 \quad y_{13} = 0 \quad y_{14} = 0\\
&y_{21} = 1 \quad y_{22} = 1 \quad\;\;\, y_{23} = 2 \quad y_{24} = 1
\end{aligned} \tag{63}$$

A perfect discrimination rule would be to choose class 2 if the feature $\mathbf{y}$ is greater than zero. These transformed patterns are nonoverlapping between the classes and, hence, the three-dimensional (3-D) observation vectors have been successfully mapped to one-dimensional (1-D) points with perfect discrimination. For comparison, the KLE transformation to 1-D fails to discriminate the data perfectly (45).

Bhattacharyya Distance

The calculation of error probability is a difficult task, even when the observation vectors have a normal pdf. Closed-form expressions for probability of error exist only for trivial, uninteresting situations. Often, the best we can hope for is a closed-form expression of some upper bound of error probability. The Bhattacharyya distance is closely tied to the probability of error as an upper bound on the Bayes error for normally distributed classes (46). For normal pdfs, the *Bhattacharyya distance* between class ω_1 and ω_2, also referred to as $\mu(1/2)$, is

$$d_{\mathrm{B}}^2 = \frac{1}{2}\ln\frac{\left|\dfrac{\mathbf{C}_i+\mathbf{C}_j}{2}\right|}{|\mathbf{C}_i|^{1/2}|\mathbf{C}_j|^{1/2}} + \frac{1}{8}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\left(\frac{\mathbf{C}_i+\mathbf{C}_j}{2}\right)^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j) \tag{64}$$

The Bhattacharyya distance directly compares the estimated mean vector and covariance matrix of the test segment with those of the target speaker. If inclusion of the test covariance in the metric is useful, Bhattacharyya distance will outperform Mahalanobis distance. Neglecting scaling, the second term is the Mahalanobis distance using an average covariance matrix. As will be shown later, if the Mahalanobis distance using an average covariance matrix performs poorly, a different pair of scale factors can yield better discrimination.

Bhattacharyya Shape. Note that Eq. (64) is the sum of two components, one is based solely upon the covariance matrices and the other involves differences between the mean vectors. These components can be characterized, respectively, as an average shape and the difference in size of the pdfs. This shape component, the *Bhattacharyya shape,* will prove very useful later on

$$d_{\mathrm{B}}' = \ln\frac{\left|\dfrac{\mathbf{C}_i+\mathbf{C}_j}{2}\right|}{|\mathbf{C}_i|^{1/2}|\mathbf{C}_j|^{1/2}} \tag{65}$$

The Bhattacharyya distance and the divergence measure have many similarities (51–54). As will be seen later, they both yield similar speaker identification performance.

The next section introduces statistical pattern matching.

PATTERN MATCHING

The pattern-matching task of speaker verification involves 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 (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

$$L = \exp(-ad) \tag{66}$$

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 (55). 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 $d(\mathbf{x}_i, \overline{\mathbf{x}})$. The model for the claimed speaker could be the centroid (mean) of a set of N training vectors

$$\overline{\mathbf{x}} = \boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \tag{67}$$

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

$$d(\mathbf{x}_i, \overline{\mathbf{x}}) = (\mathbf{x}_i - \overline{\mathbf{x}})^T\mathbf{W}(\mathbf{x}_i - \overline{\mathbf{x}}) \tag{68}$$

where $\mathbf{W}$ is a weighting matrix. If $\mathbf{W}$ is an identity matrix, the distance is *Euclidean;* if $\mathbf{W}$ is the inverse covariance matrix corresponding to mean $\overline{\mathbf{x}}$, then this is the *Mahalanobis distance*, as shown in Eq. (22). 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 (44).

Dynamic Time Warping. The most popular method to compensate for speaking-rate variability in template-based systems is known as DTW (56). 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

$$z = \sum_{i=1}^{M} d(\mathbf{x}_i, \overline{\mathbf{x}}_{j(i)}) \tag{69}$$

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

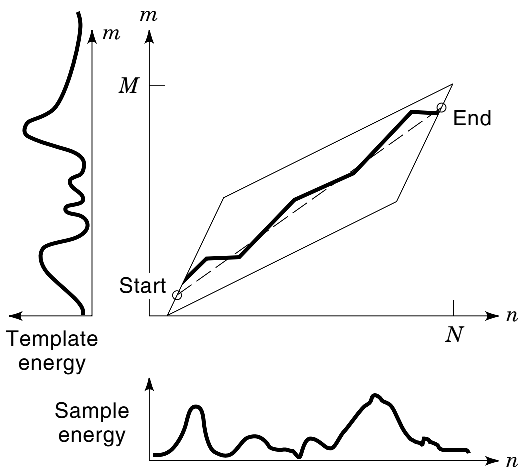

Figure 14. DTW of two energy signals.

tract. Figure 14 shows what a warp path looks like when the energies of the two speech signals are used as warp features.

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 (56), 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 (55). A VQ code book is designed by standard clustering procedures for each enrolled speaker using his training data, usually based upon reading a specific text. The pattern match score is the distance between an input vector and the minimum distance codeword in the VQ code book C. The match score for L frames of speech is

$$z = \sum_{j=1}^{L} \min_{\overline{\mathbf{x}} \in \mathbf{C}} \{d(\mathbf{x}_j, \overline{\mathbf{x}})\} \tag{70}$$

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 new method combining strengths of the DTW and VQ methods is called nearest neighbors (NN) (30,57). 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 Fig. 15, the interframe distance matrix is computed by measuring the distance between test session frames (the input) and the claimant's enrollment session frames (stored). 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 Fig. 15, 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 (30) as described in the section entitled "A New Speaker Recognition System." 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 Fig. 21.

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 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 be either 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. Unbiased estimates for the parameters of this model, the mean $\boldsymbol{\mu}$ and the covariance $\mathbf{C}$, are given by Eqs. (24) and (25), respectively. In this case, the probability that an observed feature vector $\mathbf{x}_i$ was generated by the model is

$$p(\mathbf{x}_i|\text{model}) = (2\pi)^{-k/2}|\mathbf{C}|^{-1/2} \exp\{-\tfrac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\} \tag{71}$$

Hence, $p(\mathbf{x}_i|\text{model})$ is the match score. If nothing is known about the true densities, then 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;

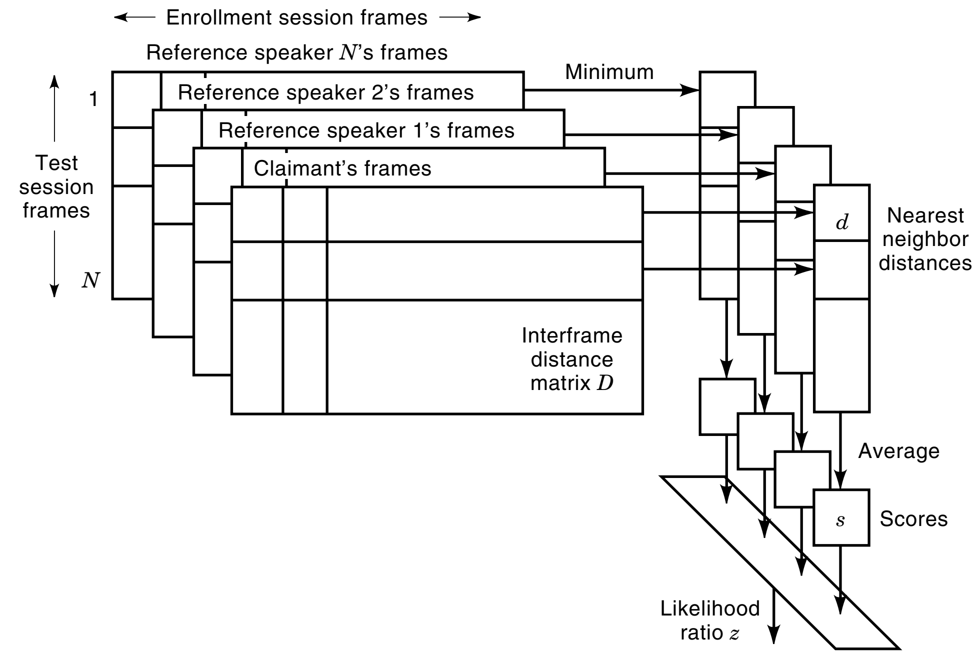

Figure 15. Nearest neighbor method.

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 be viewed only through another set of stochastic processes that produce the sequence of observations (42). 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 Fig. 16. The probability that a sequence of speech frames was generated by this model is found by using Baum–Welch decoding (58,59). This likelihood is the score for L frames of input speech given the model

$$p(\mathbf{x}(1;L)|\text{model}) = \sum_{\substack{\text{all state}\\ \text{sequences}}} \prod_{i=1}^{L} p(\mathbf{x}_i|s_i)p(s_i|s_{i-1}) \qquad (72)$$

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 (21) and more recently to outperform conventional methods in text-dependent testing (23).

Classification methods and statistical decision theory complete the system presentation and are presented in the following section.

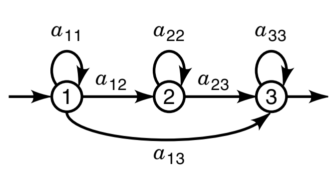

Figure 16. An example of a three-state HMM.

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). 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 (60).

Hypothesis Testing

Given a match score, the binary choice ASV classification problem involves choosing between two hypotheses: that the user is the claimed speaker or that he is not the claimed speaker (an impostor). Let H_0 be the hypothesis that the user is an impostor, and let H_1 be the hypothesis that the user is, indeed, the claimed speaker. As shown in Fig. 17, the match scores of the observations form two different pdfs according to whether the user is the claimed speaker or an impostor.

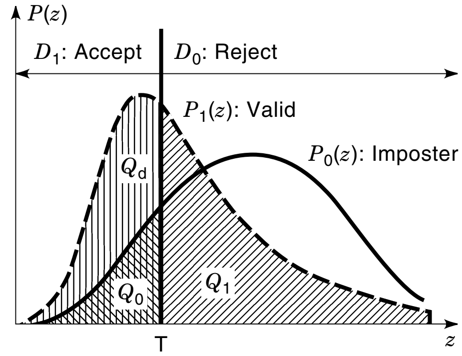

Figure 17. Valid and imposter densities.

Table 5. Probability Terms and Definitions

Performance Probabilities	Decision D	Hypothesis H	Name of Probability	Decision Result	
Q_0	1	0	Size of test "significance"	Type I error	False acceptance or alarm
Q_1	0	1		Type II error	False rejection
$Q_d = 1 - Q_1$	1	1	Power of test		True acceptance
$1 - Q_0$	0	0			True rejection

The names of the probability areas in Fig. 17 are given in Table 5. To find a given performance probability area, the hypothesis determines over which pdf to integrate, and the threshold determines which decision region forms the limits of integration.

Let $p(z|H_0)$ be the conditional density function of the observation score z generated by speakers other than the claimed speaker and likewise $p(z|H_1)$ for the claimed speaker. If the true conditional score densities for the claimed speaker and the other speakers are known, then the Bayes test with equal misclassification costs, for speaker A is based upon the likelihood ratio for speaker A, $\lambda_A(z)$ (46)

$$\lambda_A(z) \equiv \frac{p_A(z|H_0)}{p_A(z|H_1)} \tag{73}$$

Figure 18 shows an example of two score pdfs. The probability of error, which is minimized by Bayes' decision rule, is determined by the amount of overlap in the two pdfs. The smaller the overlap between the two pdfs, the smaller the probability of error. The overlap in two Gaussian pdfs with means μ_0 and μ_1 and equal variance σ can be measured by the F-ratio

$$F = \frac{(\mu_0 - \mu_1)^2}{\sigma^2} \tag{74}$$

If the true conditional score densities for the claimed speaker and other speakers are unknown, the two pdfs can be estimated from sample experimental outcomes. The conditional pdf given true speaker A, $p_A(z|H_1)$ is estimated from the speaker's own scores using his model. The conditional pdf for impostors, $p_A(z|H_0)$, is estimated from other speakers' scores using speaker A's model.

Now that the likelihood ratio for speaker A, $\lambda_A(z)$ can be determined, the classification problem can be stated as choosing a threshold T so that the decision rule is

$$\text{if } \lambda_A(z) \begin{cases} \geq T, \text{choose } H_0 \\ < T, \text{choose } H_1 \end{cases} \tag{75}$$

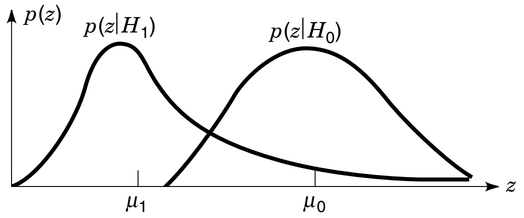

Figure 18. An example of score densities.

The threshold T can be determined by (1) setting T equal to an estimate of p_1/p_0 to approximate minimum error performance, where p_0 and p_1 are the a priori probabilities that the user is an impostor and that the user is the true speaker, respectively; (2) choosing T to satisfy a fixed FA or FR criterion (Neyman–Pearson); or (3) varying T to find different FA/FR ratios and choosing T to give the desired FA/FR ratio. With cautious constraints, T could be made speaker specific, speaker adaptive, and/or risk adaptive (e.g., break-ins may be more likely at night).

Receiver Operating Characteristic

Because either of the two types of errors can be reduced at the expense of an increase in the other, a measure of overall system performance must specify the levels of both types of errors. The tradeoff between FA and FR is a function of the decision threshold. This is depicted in the ROC curve, which plots probability of FA versus probability of FR (or FA rate versus FR rate). For example, Fig. 19 shows a hypothetical family of ROCs plotted on a log–log scale. The line of equal error probability is shown as a dotted diagonal line. The family of lines at $-45°$ represents systems with different FA·FR products, with better systems being closer to the origin. For any particular system, the ROC is traversed by changing the threshold of acceptance for the likelihood ratio. The straight line ROCs in Fig. 19 indicate that the product of the probability of FA and the probability of FR is a constant for this hypothetical system (this is not true in general) and is equal to the square of what is referred to as the equal error rate (EER). The EER is the value for which the false acceptance errors and false rejection errors are equal.

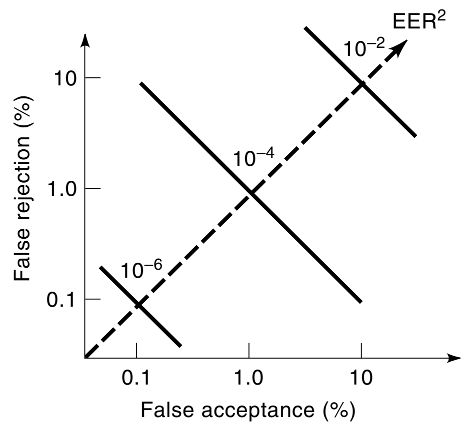

Figure 19. Hypothetical ROCs.

A NEW SPEAKER RECOGNITION SYSTEM

A simple speaker recognition system was constructed to evaluate the effectiveness of the LP-based features and information theoretic measures presented in this article. The basic building blocks needed are (1) signal acquisition, (2) feature extraction and selection, (3) pattern matching, and (4) decision criterion. The signal acquisition stage in Fig. 20 is shown for completeness; however, it is unnecessary here because the speech signal is already available in digital form from the YOHO CD-ROM. As shown in Fig. 20, the feature extraction begins with an LP analysis, followed by transformation to log area ratios [Eq. (15)], LSP frequencies [zeroes of Eq. (17)], and LP cepstra (40). The LP coefficients are estimated on unpreemphasized speech sampled at 8 kHz every 10 ms using a 10th-order autocorrelation analysis method with 20 ms overlapping Hamming windows and 15 Hz bandwidth expansion. The bandwidth expansion operation replaces the LP analysis predictor coefficients a_k by $a_k\gamma^k$, where $\gamma = 0.994$ for a 15 Hz expansion. This broadens the formant bandwidths by shifting the poles radially toward the origin in the z plane by the weighting factor γ for $0 < \gamma < 1$. This LP analysis is used in Federal Standard 1016 speech coding (41). Thus, this system is applicable to remote speaker recognition via digital speech coding.

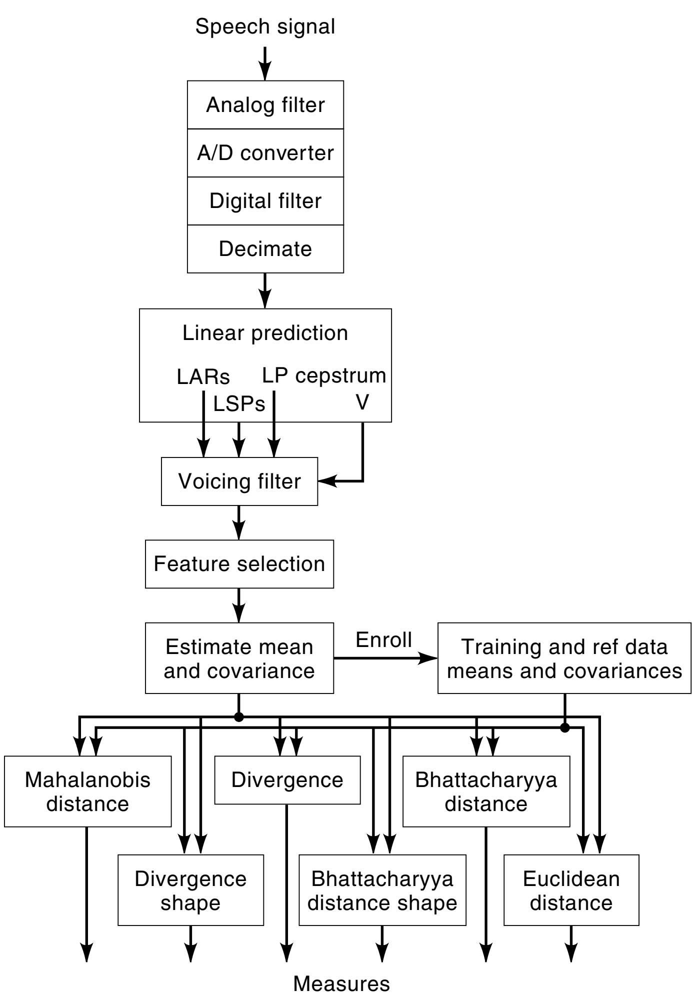

Figure 20. New speaker recognition system.

Table 6. Known Wolves and Sheep of the DTW System

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%)

As shown in Fig. 20, feature selection consists of keeping only voiced features (to reduce the effects of acoustic noise and comply with LP modeling assumptions) and forms vectors consisting of one or more of the extracted features. For example, if 10 dimensional LARs and 10-dimensional LP cepstra are selected, the resultant feature vector is their 20-dimensional concatenation, and it is used only if the frame is voiced.

During training, each speaker's mean vector [Eq. (67)] and covariance matrix [Eq. (23)] are computed and stored as a model. During testing, the recursive mean [Eq. (24)] and recursive covariance [Eq. (25)] are computed and compared with the stored models. Using the recursive estimates allows the comparisons to occur as the speech sample is being taken so that early recognition decisions can be made. The mean vector and covariance matrix used to model each speaker can be compactly represented. For the shape measures, only the covariance matrix is needed. For a 10-dimensional feature (e.g., the LSPs from a 10th-order LP analysis), each speaker is represented by the covariance matrix of his 10 LSP frequencies. Because of symmetry, a covariance matrix can be uniquely represented by its upper (or lower) triangular section. Exploiting this symmetry, a person's 10×10 covariance matrix can be represented with only 55 elements, thus allowing for very compact speaker models.

Various measures are computed to be evaluated in combination with various features. The following measures are computed for pattern matching: the divergence shape [Eq. (41)], Bhattacharyya shape [Eq. (65)], Bhattacharyya distance [Eq. (64)], divergence measure [Eq. (40)], Mahalanobis distance [Eq. (22)], and Euclidean distance [Eq. (68)].

Last, the decision criterion is to choose the closest speaker according to the selected feature and measure (this criterion suffices for evaluating features and measures, but it is incomplete for open-set conditions). For most real-world applications, where open set impostors exist, thresholding the match score to ensure some degree of closeness is necessary before making a recognition decision. Threshold determination should account for the costs of different types of errors the system can commit (e.g., a false acceptance error might be more costly than a false rejection error) and the probabilities of those errors occurring, which might vary (e.g., attacks might be more likely at night than during the day).

The LSP features used with the divergence shape measure is shown to have strong speaker discriminatory power in the

Table 7. Wolf and Sheep Sexual Characteristics

19 FA Errors Across 9,300 Impostor Trials		
Number of FA Errors	Wolf Sex	Sheep Sex
15	Males	Males
1	Female	Female
3	1 male	3 females

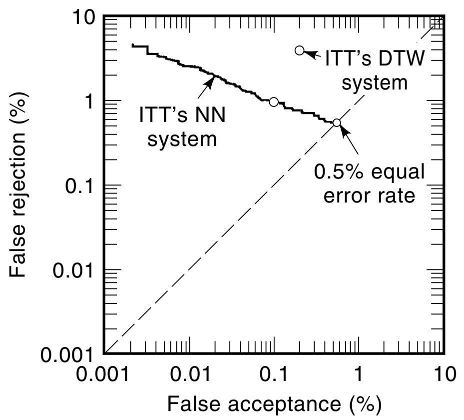

Figure 21. Receiver operating characteristics.

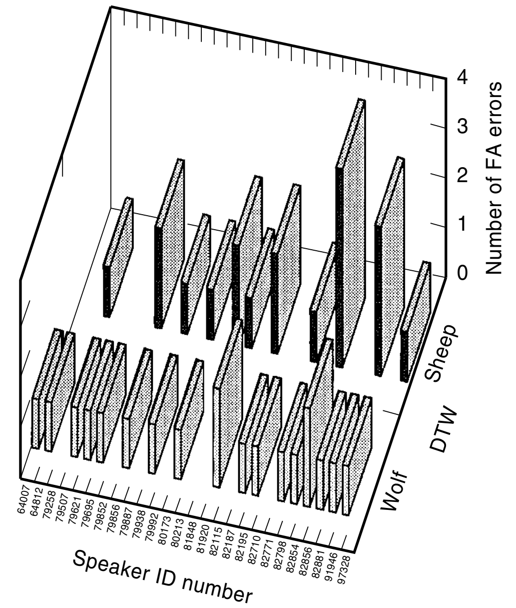

Figure 22. Speaker versus FA errors for the DTW system's wolves and sheep.

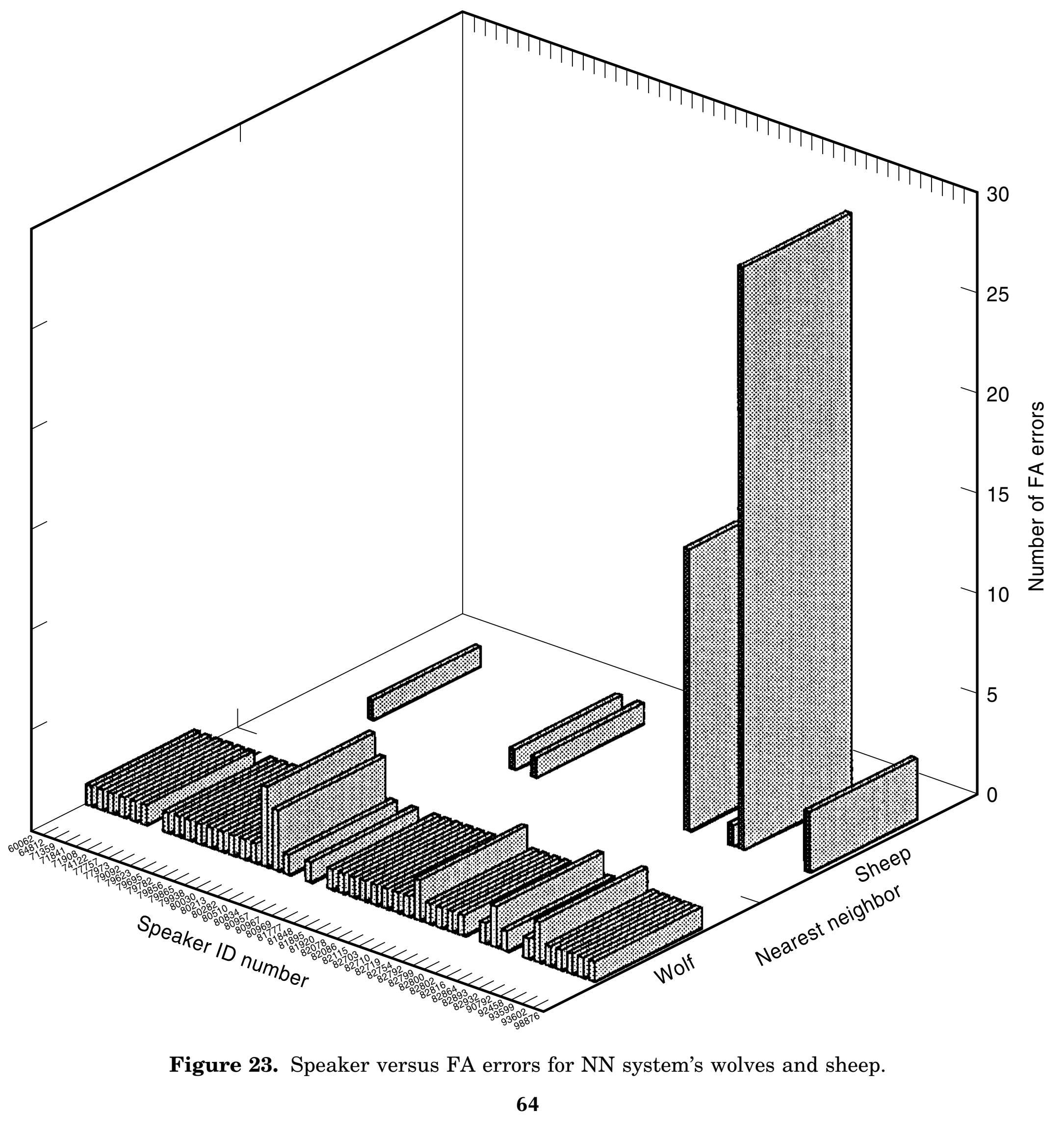

Figure 23. Speaker versus FA errors for NN system's wolves and sheep.

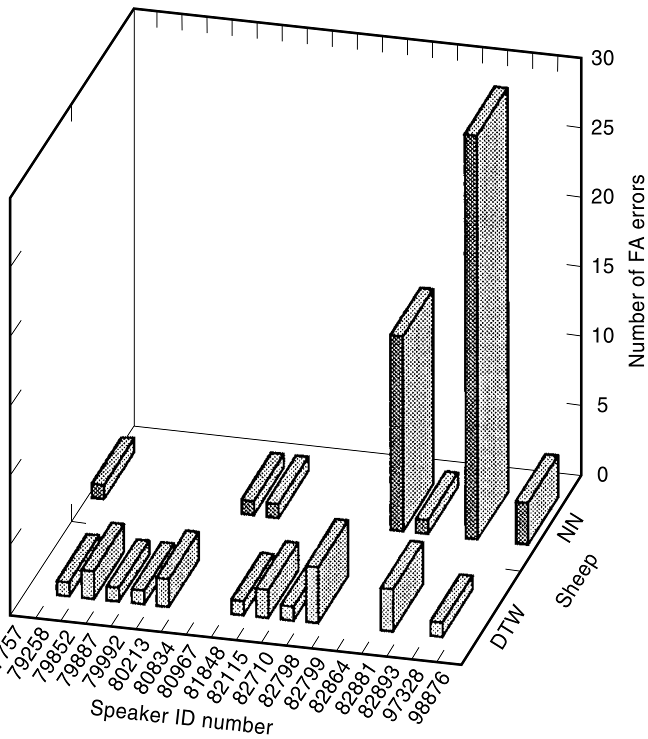

Figure 24. Speaker versus FA errors for DTW and NN systems' sheep.

following section. The LSP and LP cepstral features are also found to be powerful when used with the divergence measures and Bhattacharyya distances.

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. The full YOHO database has 10 test sessions for each of 186 subjects. For only one test session, there are

$$\binom{186}{2} = 17,205$$

pairwise combinations. Because of computational limitations, not all pairwise combinations for all 10 test sessions were tested. Thus, the simulated imposter testing drew randomly across the 10 test sessions. 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 (12,30).

DTW System

The DTW ASV system tested here was created by Higgins et al. (10). 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 (10).

Cross-speaker testing (causal 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 6 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 vary narrow portion of the population, especially if two errors are required to designate a person as a wolf or sheep. The difficulty in acquiring enough data to represent the wolf and sheep populations adequately makes it challenging to study these errors.

From the 9,300 trials, there were 19 FA errors for the DTW system. Table 7 shows that these 19 pairs of wolves and sheep have interesting sexual 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 database provides at least 19 pairs of wolves and sheep under the DTW system for further investigation. It should be noted that because of computational limitations, not all possible wolf and sheep combinations have been tested. Even with this large database, relatively few wolves and sheep have been discovered to date.

ROC of DTW and NN Systems

Figure 21 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 the following sections, the two systems commit different errors.

Wolves and Sheep

FA errors due to individual wolves and sheep are shown in the 3-D histogram plots of Figs. 22–25. Figure 22 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.

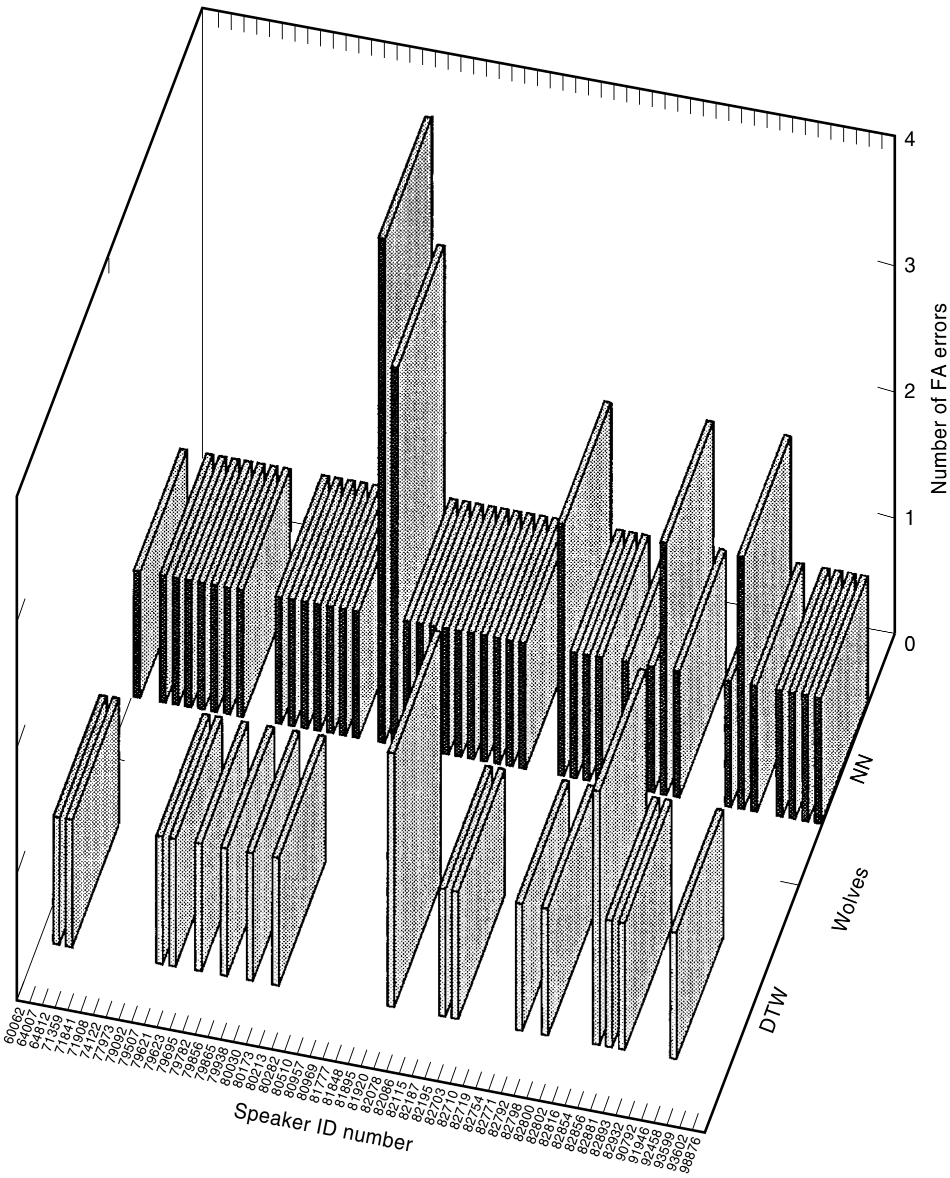

Figure 25. Speaker versus FA errors for DTW and NN systems' wolves.

The DTW system rarely has the same speaker as both a wolf and a sheep (there are only two exceptions in these 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 23 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.

Now we'll investigate the relations between the NN and DTW systems. Figure 24 shows the sheep of the NN and DTW systems. It should be noted from Fig. 24 that the two sheep who dominate the FA errors of the NN system were not found to be sheep in the DTW system. This suggests the potential for making a significant performance improvement by combining the systems.

Figure 25 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. In fact, a speaker detection system consisting of eight combined systems has been demonstrated recently (11).

Figure 26 shows the number of FA errors that occur for various test sessions of the NN system. The figure clearly shows that a couple of sessions (namely, numbers 880 and 1858) have an excessive number of FA errors. Upon listening to sessions 880 and 1858, it sounds like these sessions have more boominess than the other test (and enrollment) sessions. The acoustic environment might have changed during these problem sessions.

Wolves and sheep come in pairs. Figure 27 shows the DTW system's wolf and sheep pairings for the YOHO database. It should be noted that under the DTW system, speaker 82798 is a particularly vulnerable sheep with respect to wolves 81920, 82866, and 79866. These speakers, in addition to the

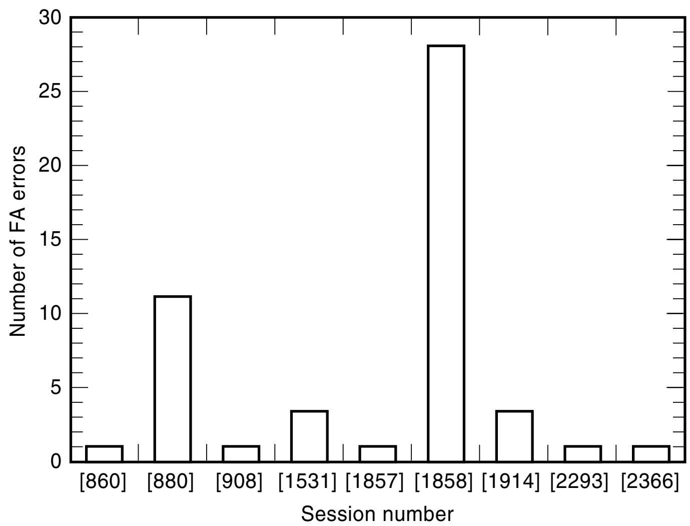

Figure 26. FA errors versus session number for NN system.

others shown in Fig. 27, will be of prime interest in the following experiments.

New Speaker Recognition System

The new speaker recognition system, described earlier, was evaluated in close-set speaker identification testing. Speaker identification experiments using 44 and 43 speaker subsets of the YOHO database were performed. In the 44 person test from the YOHO database, each speaker is compared to a different session of himself and to 2 sessions of 43 other speakers using 80 s of speech for training and a separate 80 s of speech for testing.

In the mesh plots of Figs. 28–31, each of the 44 people are shown along the i and j axes; the i axis represents speech collected from session 1 versus the j axes, with speech collected from session 2. Thus, there are 44^2 measures, each represented by a point on the mesh. The z axis is the reciprocal of the measure indicated in the figure's caption using LSP features. Thus, "close" speakers will cause a peak along the z axis. The ideal structure, representing perfect speaker identification, would be a prominent diagonal such that $a_{ii} > a_{ij}$ $\forall$ $i \neq j$.

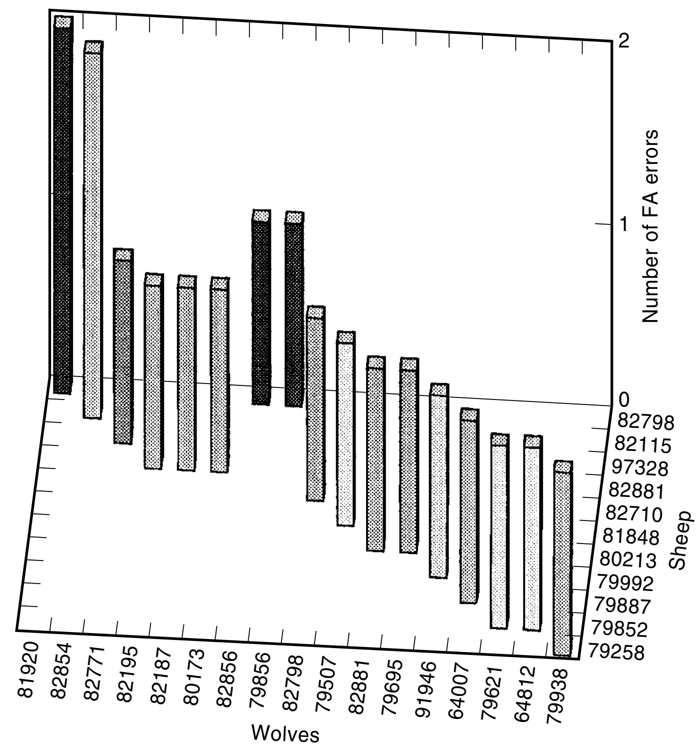

Figure 27. Wolf and sheep pairings of the DTW system.

Notice the nearly ideal prominent diagonal structure in Fig. 28 provided by the LSP divergence shape; thus, its discrimination power is strong. The single confusion error made by the LSP divergence shape, shown by an arrow in Fig. 28, is between session 1 of speaker 59771 and session 2 of speaker 79082. It is interesting to note that this is not one of the DTW system's pairs of wolves and sheep, as sheep in Fig. 27. It is also interesting to note that this same error occurs in all the LSP-based divergence and Bhattacharyya distance systems, as shown by a peak at the same location as the arrow in Fig. 28 in each of the mesh plots in Figs. 29–31.

Notice the similarity in structure between the mesh plots of the LSP Bhattacharyya shape shown in Fig. 29 and the LSP divergence shape. Not only do these measures perform similarly well, but the measures also appear to be related.

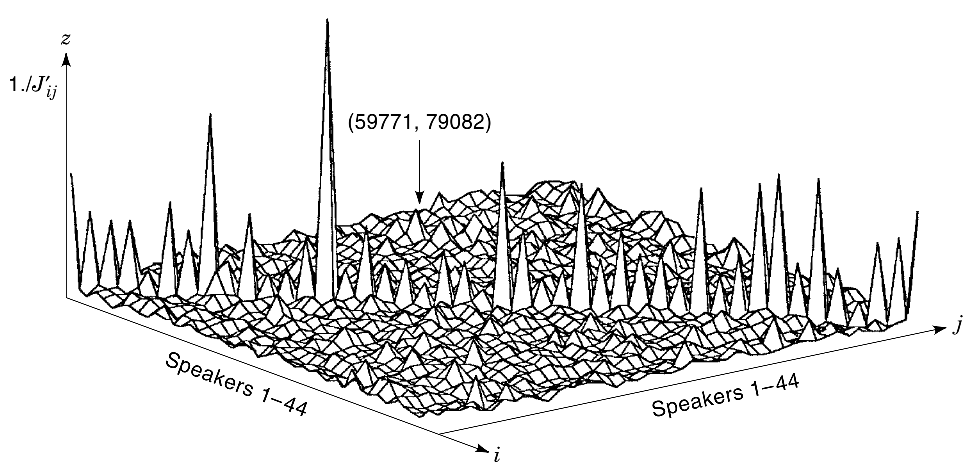

Figure 28. LSP divergence shape (1 error).

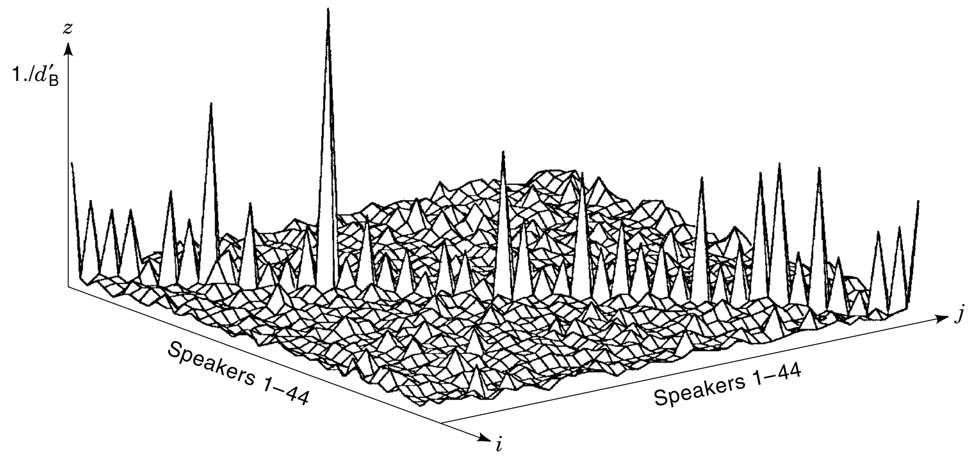

Figure 29. LSP Bhattacharyya shape (2 errors).

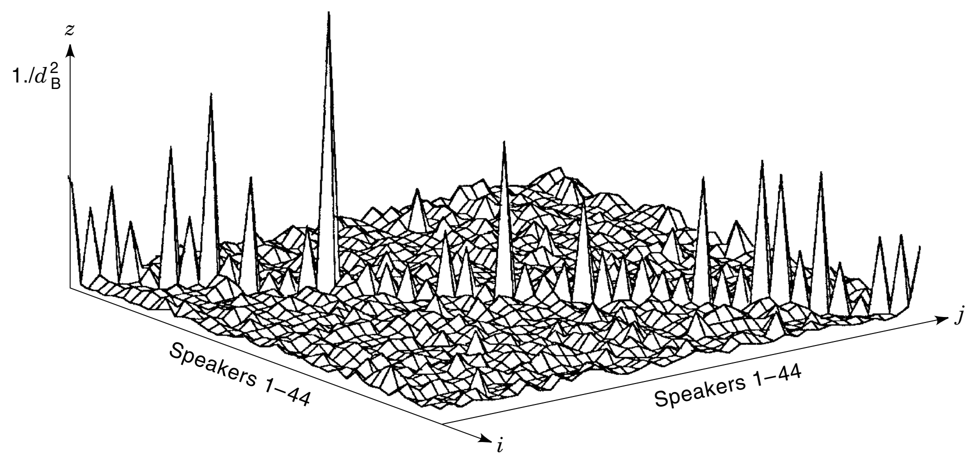

Figure 30. LSP Bhattacharyya distance (4 errors).

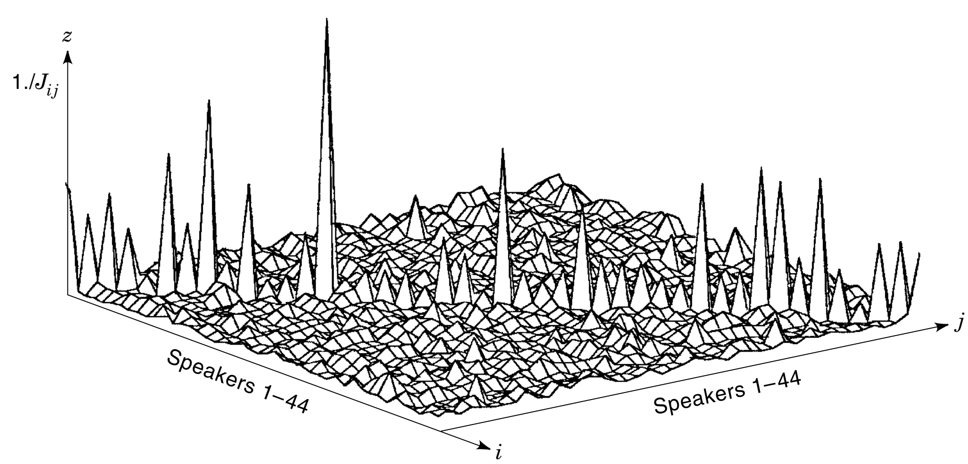

Figure 31. LSP divergence measure (3 errors).

Table 8. Confusions Using Various Features and Measures

	LSP	LP Cepstrum	LAR
Divergence shape	0.05%	0.15%	
Bhattacharyya shape	0.10%	0.10%	
Bhattacharyya distance	0.21%	0.10%	
Divergence measure	0.15%	0.21%	0.52%
Mahalanobis distance		1.08%	
Euclidean distance		1.96%	

Note the slight degradation in performance of the LSP Bhattacharyya distance in Fig. 30 versus the LSP Bhattacharyya shape. The inclusion of the means in the Bhattacharyya distance degraded its performance. This discovery provided the insight toward the development of the shape measures.

Note the degraded performance of the LSP divergence measure in Fig. 31 relative to the divergence shape. Again, inclusion of the means degraded the performance.

The power of using the LSP features in these measures is shown by the prominent diagonal structure in the previous figures.

The results are summarized in Table 8, with additional identification experiments performed on the same data. Out of the 1936 measures, Euclidean distance commits 38 confusion errors (1.96% error) and Mahalanobis distance makes 21 confusion errors (1.08% error) when using LP cepstrum combined with LAR features. The LSP divergence shape performs the best among these experiments with only 1 confusion error (0.05%). A single confusion error across the 88 identification tests corresponds to a 1.1% closed-set speaker identification error rate. One might conclude from these results that the means of the features tested tend to be unreliable, whereas the variances and covariances in the features have reliable discrimination power. In fact, the author was led to the divergence shape and Bhattacharyya shape (removing the means) by the mediocre performance of the Euclidean and Mahalanobis distances.

The simple LSP divergence shape is shown to have speaker-discriminatory power. The LSP and LP cepstral features were found to be powerful in the divergence measures and Bhattacharyya distances. The LSP divergence shape performs the best among these tests with only one confusion error (0.05%); however, a larger test would be needed to claim that this is significantly better than the Bhattacharyya-distance-based results.

Last, we conclude by reviewing the problem at hand and summarizing the major concepts of this article.

SUMMARY

Automatic speaker recognition is the use of a machine to recognize a person from a spoken phrase. Speaker recognition systems can be used in two modes: to *identify* a particular person or to *verify* a person's claimed identity. The basics of speaker recognition have been covered, and simple features and measures for speaker recognition were presented and compared with traditional ones using speaker discrimination criterion. The scope of this work is limited to speech collected from cooperative users in real-world office environments and without adverse microphone or channel impairments.

A new speaker recognition system that uses an information-theoretic shape measure and LSP frequency features to discriminate between speakers was presented. This measure, the *divergence shape,* can be interpreted geometrically as the shape of an information-theoretic measure called divergence. The LSP frequencies were found to be effective features in this divergence shape measure. A speaker-identification test yielded 98.9% correct closed-set speaker identification using cooperative speakers with high-quality telephone-bandwidth speech collected in real-world office environments under a constrained grammar across 44 and 43 speaker subsets of the YOHO corpus, with 80 s of speech for training and testing. The new speaker recognition system presented here is practical to implement in software on a modest personal computer.

BIBLIOGRAPHY

1. B. S. Atal, Automatic recognition of speakers from their voices, *Proc. IEEE,* **64**: 460–475, 1976.
2. G. R. Doddington, Speaker recognition—Identifying people by their voices, *Proc. IEEE,* **73**: 1651–1664, 1985.
3. S. Furui, Speaker-dependent-feature extraction, recognition and processing techniques, *Speech Commun.,* **10**: 505–520, 1991.
4. D. O'Shaughnessy, *Speech Communication, Human and Machine: Digital Signal Processing,* Reading, MA: Addison-Wesley, 1987.
5. A. Rosenberg, Automatic speaker verification: A review, *Proc. IEEE,* **64**: 475–487, 1976.
6. A. E. Rosenberg and F. K. Soong, Recent research in automatic speaker recognition, in S. Furui and M. M. Sondhi (eds.), *Advances in Speech Signal Processing,* New York: Dekker, 1992, pp. 701–738.
7. A. Sutherland and M. Jack, Speaker verification, in M. Jack and J. Laver (eds.), *Aspects of Speech Technology,* Edinburgh, UK: Edinburgh Univ. Press, 1988, pp. 185–215.
8. R. Mammone, X. Zhang, and R. Ramachandran, Robust speaker recognition—A feature-based approach, *IEEE Signal Process. Mag.,* **13** (5): 58–71, 1996.
9. D. Reynolds and R. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models, *IEEE Trans. Speech Audio Process.,* **3**: 72–83, 1995.
10. A. Higgins, L. Bahler, and J. Porter, Speaker verification using randomized phrase prompting, *Digital Signal Process.,* **1** (2): 89–106, 1991.
11. A. Martin and M. Przybocki, 1997 speaker recognition evaluation, in A. Martin (ed.), *Speaker Recognition Workshop,* Linthicum Heights, MD: Maritime Inst. of Technol., 1997, Sect. 2. Available ftp://jaguar.ncsl.nist.gov/speaker/ and http://www.nist.gov/itl/div894/894.01/
12. J. Campbell, Testing with the YOHO CD-ROM voice verification corpus, *Int. Conf. Acoust., Speech, Signal Process.,* Detroit, MI, 1995, pp. 341–344. Available http://www.biometrics.org/
13. B. S. Atal, Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, *J. Acoust. Soc. Amer.,* **55**: 1304–1312, 1974.
14. J. D. Markel and S. B. Davis, Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base, *IEEE Trans. Acoust. Speech Signal Process.,* **ASSP-27**: 74–82, 1979.
15. S. Furui, Cepstral analysis technique for automatic speaker verification, *IEEE Trans. Acoust. Speech Signal Process.,* **ASSP-29**: 254–272, 1981.

16. R. Schwartz, S. Roucos, and M. Berouti, The application of probability density estimation to text independent speaker identification, *Int. Conf. Acoust., Speech, Signal Process.,* Paris, 1982, pp. 1649–1652.
17. K. P. Li and E. H. Wrench, Jr., Text-independent speaker recognition with short utterances, *Int. Conf. Acoust., Speech, Signal Process.,* Boston, 1983, pp. 555–558.
18. F. Soong et al., A vector quantization approach to speaker recognition, *IEEE, Int. Conf. Acoust., Speech, Signal Process.,* Tampa, Florida, 1985, pp. 387–390.
19. A. L. Higgins and R. E. Wohlford, A new method of text-independent speaker recognition, *Int. Conf. Acoust., Speech, Signal Process.,* Tokyo, 1986, pp. 869–872.
20. J. Attili, M. Savic, and J. Campbell, A TMS32020-based real time, text-independent, automatic speaker verification system, *Int. Conf. Acoust., Speech, Signal Process.,* New York, 1988, pp. 599–602.
21. N. Z. Tishby, On the application of mixture AR hidden Markov models to text independent speaker recognition, *IEEE Trans. Acoust., Speech, Signal Process.,* **39**: 563–570, 1991.
22. D. Reynolds, Speaker identification and verification using Gaussian mixture speaker models, *Speech Commun.,* **17** (1–2): 91–108, 1995.
23. D. Reynolds and B. Carlson, Text-dependent speaker verification using decoupled and integrated speaker and speech recognizers, *EUROSPEECH ESCA,* Madrid, 1995, pp. 647–650.
24. C. Che and Q. Lin, Speaker recognition using HMM with experiments on the YOHO database, *EUROSPEECH, ESCA,* Madrid, 1995, pp. 625–628.
25. J. Colombi et al., Cohort selection and word grammar effects for speaker recognition, *IEEE, Int. Conf. Acoust., Speech, Signal Process.,* Atlanta, GA, 1996, pp. 85–88.
26. D. Reynolds, M.I.T. Lincoln Laboratory site presentation, in A. Martin (ed.), *Speaker Recognition Workshop,* Linthicum Heights, MD: Maritime Inst. of Technol., 1996, Sect. 5. Available ftp:jaguar.ncsl.nist.gov/speaker/ and http://www.nist.gov/itl/div894/894.01/
27. A. E. Rosenberg et al., The use of cohort normalized scores for speaker verification, *Int. Conf. Spoken Lang. Process.,* Banff, Univ. of Alberta, 1992, pp. 599–602.
28. H. Gish and M. Schmidt, Text-independent speaker identification, *IEEE Signal Process. Mag.,* **11** (4): 18–32, 1994.
29. G. Papcun, Commensurability among biometric systems: How to know when three apples probably equals seven oranges, *Proc. Biometric Consortium, 9th Meet.,* Crystal City, VA, 1997. Available http://www.biometrics.org/
30. A. Higgins, YOHO speaker verification, *Speech Res. Symp.,* Baltimore, MD, 1990.
31. J. Flanagan, *Speech Analysis Synthesis and Perception,* 2nd ed., Berlin: Springer-Verlag, 1972.
32. T. Parsons, Voice and speech processing, in S. Director (ed.), *Communications and Signal Processing,* New York: McGraw-Hill, 1987.
33. A. Pentz, *Speech Science (SPATH 4313) Class Notes,* Stillwater: Oklahoma State Univ., 1990.
34. D. Plumpe, Modeling of the glottal flow derivative waveform with application to speaker identification, M.S. thesis, Massachusetts Inst. of Technol., Cambridge, MA, 1997.
35. J. Makhoul, Linear prediction: A tutorial review, *Proc. IEEE,* **63**: 561–580, 1975.
36. F. J. Harris, On the use of windows for harmonic analysis with the DFT, *Proc. IEEE,* **66**: 51–83, 1978.
37. F. Itakura, Line spectrum representation of linear predictive coefficients, *Trans. Comm. on Speech Res., Acoust. Soc. Jpn.,* **S75**: 34, 1975.
38. S. Saito and K. Nakata, *Fundamentals of Speech Signal Processing,* Tokyo: Academic Press, 1985.
39. G. Kang and L. Fransen, *Low Bit Speech Encoder Based on Line-Spectrum-Frequency,* NRL Rep. 8857, Washington, DC: NRL, 1985.
40. L. Rabiner and R. Schafer, Digital processing of speech signals, in A. Oppenheim (ed.), *Signal Processing,* Englewood Cliffs, NJ: Prentice-Hall, 1978.
41. J. P. Campbell, Jr., T. E. Tremain, and V. C. Welch, The Federal Standard 1016 4800 bps CELP voice coder, *Digital Signal Process.,* **1** (3): 145–155, 1991.
42. L. Rabiner and B.-H. Juang, Fundamentals of speech recognition, in A. Oppenheim (ed.), *Signal Processing,* Englewood Cliffs, NJ Prentice-Hall, 1993.
43. R. Gnanadesikan and J. R. Kettenring, Discriminant analysis and clustering, *Stat. Sci.,* **4** (1): 34–69, 1989.
44. R. Duda and P. Hart, *Pattern Classification and Scene Analysis,* New York: Wiley, 1973.
45. J. Tou and R. Gonzalez, Pattern recognition principles, in R. Kalaba (ed.), *Applied Mathematics and Computation,* Reading, MA: Addison-Wesley, 1974.
46. K. Fukunaga, Introduction to statistical pattern recognition, in W. Rheinboldt and D. Siewiorek (eds.), *Computer Science and Scientific Computing,* 2nd ed., San Diego, CA: Academic Press, 1990.
47. S. Kullback, *Information Theory and Statistics,* New York: Dover, 1968.
48. R. E. Blahut, *Principles and Practice of Information Theory: Electrical and Computer Engineering,* Reading, MA: Addison-Wesley, 1987.
49. S. Kullback and R. Leibler, On information and sufficiency, *Ann. Math. Stat.,* **22**: 79–86, 1951.
50. J. Tou and P. Heydorn, Some approaches to optimum feature extraction, in J. Tou (ed.), *Computer and Information Sciences-II,* New York: Academic Press, 1967, pp. 57–89.
51. M. Basseville, Distance measures for signal processing and pattern recognition, *Signal Process.,* **18**: 349–369, 1989.
52. P. A. Devijver, On a new class of bounds on Bayes risk in multihypothesis pattern recognition, *IEEE Trans. Comput.,* **C-23**: 70–80, 1974.
53. T. Kailath, The divergence and Bhattacharyya distance measures in signal selection, *IEEE Trans. Commun. Technol.,* **15**: 52–60, 1967.
54. Y.-T. Lee, Information—Theoretic distortion measures for speech recognition, *IEEE Trans. Acoust. Speech Signal Process.,* **39**: 330–335, 1991.
55. F. K. Soong et al., A vector quantization approach to speaker recognition, *AT&T Tech. J.,* **66** (2): 14–26, 1987.
56. H. Sakoe and S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, *IEEE Trans. Acoust. Speech Signal Process.,* **ASSP-26**: 43–49, 1978.
57. A. Higgins, L. Bhaler, and J. Porter, Voice identification using nearest neighbor distance measure, *Int. Conf. Acoust., Speech, Signal Process.,* Minneapolis, MN, 1993, pp. 375–378.
58. L. Rabiner and B.-H. Juang, An introduction to hidden Markov models, *IEEE Acoust. Speech Signal Process. Mag.,* **3**(1): 4–16, 1986.
59. L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE,* **77**: 257–286, 1989.
60. A. Wald, *Sequential Analysis,* New York: Wiley, 1947.

JOSEPH P. CAMPBELL, JR.
The Johns Hopkins University

SPECIFICATION OF SOFTWARE. See FORMAL SPECIFICATION OF SOFTWARE.