Speaker Verification: A Tutorial

Jayant M. Naik

PERSONAL IDENTITY VERIFICATION IS AN essential requirement for controlling access to protected resources. Personal identity is usually claimed by presenting a unique personal possession such as a key, a badge, or a password. However, these can be lost, stolen, or counterfeited, thereby posing a threat to security. Further, a simple identity claim might be insufficient if the potential for loss is great and the penalty for false identification is severe. Hence, verification of that claimed identity is necessary. This can be attempted by examining an individual's biometric features, such as fingerprints, hand geometry, or retinal pattern, or by examining certain features derived from the individual's unique activity, such as speech or handwriting. In each case, the features are compared with the previously stored features for the person whose identity is being claimed. If this comparison is favorable, based on a decision criterion, then the claimed identity is verified.

Among these methods, identity verification based on a person's voice has special advantages for practical deployment. Speech is our most natural means of communication and, therefore, user acceptance of the system would be very high. Advances in digital signal processors and speech technology have made possible the design of fast, cost effective, high-performance speaker verification systems. These systems can be easily integrated into the ubiquitous telephone network, thereby providing access control for banking transactions by telephone, automatic telephone transactions such as voice mail and credit card verification, and remote access to computers via modems on dial-up telephone lines.

Task Definition

The task of speaker verification is a subset of the general problem of speaker recognition, which includes the task of speaker identification. Speaker identification means labeling an unknown voice as one of a set of known voices, whereas speaker verification means determining whether an unknown voice matches the known voice of a speaker whose identity is being claimed. Since the speaker verification task involves a binary comparison, its performance (measured in terms of error probabilities) is independent of population size. On the other hand, the speaker identification task involves $N + 1$ decisions for a population size of N speakers (deciding that the unknown voice is one of N known voices or none of them); hence, its performance degrades with an increasing number of users [2].

The performance of the above two tasks is further determined by the type of speech material used to claim an identity. Fixed-text systems require the recitation of a predetermined text, thereby maintaining a high degree of user cooperation, whereas free-text systems accept speech utterances of unrestricted text. In fixed-text systems, with adequate time alignment, one can make precise and reliable comparisons between two utterances of the same text, in similar phonetic environments. This is not easily accomplished in free-text systems. Hence, fixed-text systems have a much higher level of performance than free-text systems. Free-text systems often use long-term statistics of the speech signal to extract speaker-specific data and require 10–30 s of speech for training and 5–10 s of speech for verification. Fixed-text systems typically require 2–3 s of speech for training and for verification. While fixed-text systems are primarily used in access control applications where the user is cooperative and consistent from session to session, free-text systems are needed in forensic and surveillance applications where the user is not cooperative and often not aware of the task. Fixed-text systems frequently employ some randomization such as random sequence of digits or predefined words, as a means of preventing the use of audio recordings by determined intruders.

Personal identity is usually claimed by presenting a unique personal possession such as a key, a badge, or a password.

Feature Selection

The speech signal conveys information about the speaker in many ways. These include "high-level" features such as dialect, context, speaking style, emotional state of the speaker, etc., which are often used by human listeners to identify a person. Efforts have been made to identify the perceptual bases of speaker verification used by human listeners and to select their acoustic correlates derived from the speech signal [26]. These efforts have not been successful because of the difficulty in acquiring and quantitatively measuring the speaker-discriminating features used by humans. Furthermore, training and evaluating a speaker verification system based on these perceptual correlates is impractical. Hence, operational speaker verification systems use "low-level" parameters, such as pitch, spectral magnitudes, formant frequencies, energy pro-

0163-6804/90/0001-0042 $01.00 © 1990 IEEE

files, etc. [1] [2], which are derived from acoustic measurements of the speech signal.

The speech signal is a complex function of the speaker's physical characteristics, such as vocal source/tract dimensions, environment (e.g., background noise and transmission channel), and emotional state, such as physical and mental stress. Hence, there are large variabilities in the speech signal between speakers and, more importantly, between speech data collected from the same speaker at different times. A judicious selection of the acoustic features is crucial for the effectiveness of a speaker verification system. The features used in any speaker verification system should:

- Discriminate between speakers while being tolerant of intra-speaker variabilities
- Be easily measurable from the speech signal
- Be stable over time
- Not be susceptible to mimicry by impostors

The speech signal is a complex function of the speaker's physical characteristics, such as vocal source/tract dimensions, environment (e.g., background noise and transmission channel), and emotional state, such as physical and mental stress.

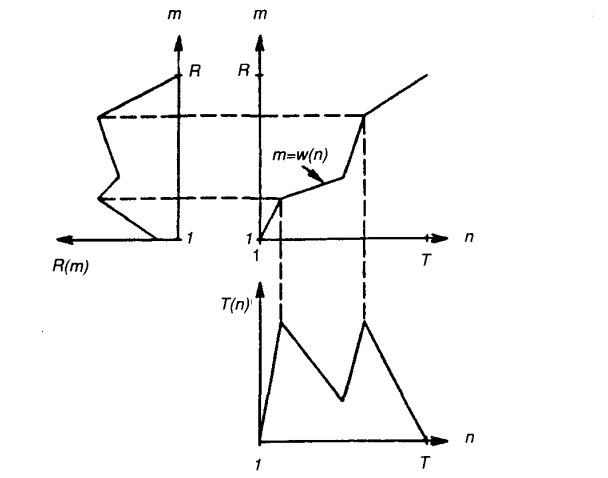
The effectiveness of various features has been studied extensively [3] [4]. Voice pitch is an obvious choice and has been employed in several speaker verification systems [5], but it is not always easy to measure, especially in noisy environments, and is easy to mimic. Stress and speech effort levels also change the pitch patterns significantly. Speech intensity, as a function of time, has been shown to be effective [10] [11]. Formant frequencies (resonant frequencies of the vocal tract) contain speaker-specific information and can be used to distinguish between speakers. The main drawback is the difficulty of measurement, but computationally efficient methods of formant frequency measurements have been proposed [12] [13]. Coarticulation during the production of a nasal sound has also been employed as a speaker-specific feature [14].

The short-term spectrum of the speech signal, defined as a function of time, frequency, and spectral magnitude, is by far the most prevalent method of representation of the speech signal [1]. Several approximations to the short-term spectrum, such as filter bank magnitudes, and Linear Prediction Coding (LPC) spectral and cepstral coefficients, are also popular [15–19]. The choice of a particular representation is determined by practical considerations, such as ease of computation, storage requirements, methods of pattern matching, susceptibility to transmission channel distortions, etc. Recent work in speaker discrimination modeling [27], which used a linear transformation of the features designed to enhance the separation between valid users and impostors, has yielded significant improvements in speaker verification performance over telephone lines.

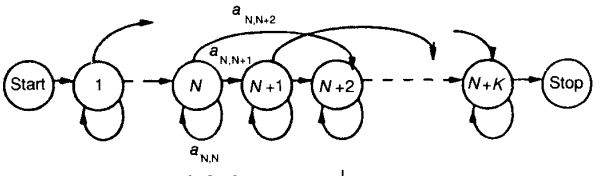
Pattern Matching

A popular method of pattern matching in speaker verification systems is based on template matching, in which the speech signal is represented in terms of a series of feature vectors that characterize the behavior of the speech signal for a particular speaker. This time-ordered set of features is called a template. A template can be a multi-word utterance, a single word, a syllable, or a phoneme. Most speaker verification systems use either a word or a multi-word sentence as a template. In a template-matching scheme, the comparison between an input utterance template and the reference template is performed by aligning the two templates at equivalent points in time. The durations of the reference and test templates will invariably be different, and some stretching or compressing will be necessary. A popular technique to accomplish this task is a dynamic programming method that uses an optimum time expansion/compression function for nonlinear time alignment, as shown in Figure 1(a). This is commonly known as Dynamic Time Warping (DTW). A distance metric, defined as a function of time, is computed between the two feature sets representing the speech data. A decision function is then derived by integrating this metric over time. This decision function is a by-product of the DTW time alignment process [28–30]. Several distance measures for comparing the speech feature sets have been studied [6]. They include the Itakura distance for comparing two sets of linear predictor coefficients and the Euclidian distance [30].

An alternative to the template-matching approach is to build probabilistic models of the speech signal that describe its time-varying characteristics. This is known as the Hidden Markov Modeling (HMM) technique, which is being used in a number of speaker verification algorithms. It is a doubly stochastic process in that it has an underlying stochastic process that is not observable (hence the term hidden), but can be observed through another stochastic process that produces a sequence of observations [31]. The HMM technique represents the model of speech production as a system that is capable of being in only a finite number of different states. Each state is



(a) Non-linear time alignment of test pattern $T(n)$ with reference pattern $R(n)$



(b) Left-to-right HMM structure.

Fig. 1. Pattern-matching techniques.

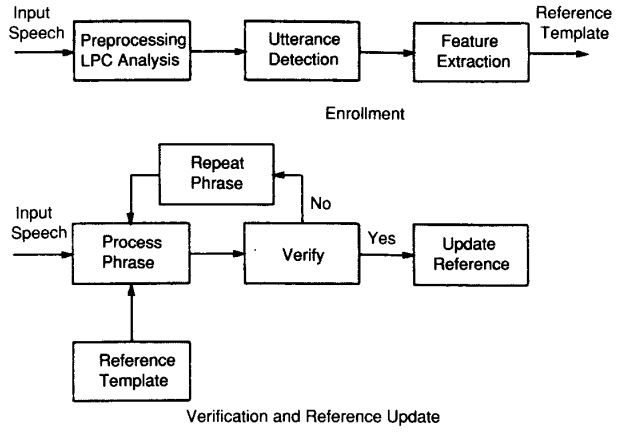


Fig. 2. Fixed-text, template-based speaker verification system stages.

capable of generating either a finite number of outputs (the Vector Quantization approach) or a continuum of outputs (the continuous distribution approach). At discrete intervals of time, the system passes from one state to another, each state producing an output. The transitions between the states are random, as are the outputs associated with each state. By allowing random transitions and outputs, the model accommodates the temporal variations of the speech signal. A popular model structure is the "left-to-right" model shown in Figure 1(b). It has a single starting state and a single final state, and state N has transitions to the same state, to state $(N + 1)$ or a "skip" transition to state $(N + 2)$. A state is defined every 20 ms. The HMM system consists of the following operations:

- Optimize the model parameters so as to best describe the observed sequence (Training)
- Given an observation sequence, choose a state sequence that is optimal according to some predefined criterion
- Given an observation sequence and the model, compute the probability of the observation sequence (Scoring)

A speaker verification system based on an HMM architecture can use speaker models derived from a multi-word sentence, a single word, or a phoneme. Typically, multi-word phrases (a string of seven to ten digits, for example) are used, and models for each individual word and for "silence" are combined at a sentence level according to a predefined sentence-level grammar [27] [32].

Speaker Verification System Design

A typical template-based, fixed-text speaker verification task consists of:

- Enrollment: Creation of a set of speech features, as a function of time, for each valid user
- Verification: Comparison of input speech with reference templates at equivalent points in time; decision based on similarity between input and reference, integrated over time
- Reference Update: Adaptation of the reference templates to accommodate changes in the valid user's speech after successful verification

The flow chart for a verification task is shown in Figure 2.

The enrollment procedure builds an initial reference template for a speaker by capturing the verification utterance. Typically, this is accomplished by examining the speech energy profile and determining the end points in time. Since this template is used for future comparisons, extreme care should be taken in generating a good enrollment template. A thorough user indoctrination is necessary to obtain a good enrollment template and to maintain consistent recitation of the voice password (verification phrase). When enrollment has to be performed without supervision, as in the case of speaker verification over the telephone network, the enrollment procedure should be secured to prevent enrollment by unauthorized persons, for example, by requiring an additional password.

A sequential strategy is often employed to improve the valid user acceptance rate. This involves the use of more than one utterance (usually two to five) and a decision strategy that sets different acceptance thresholds for each individual utterance, or combinations of two or more utterances [7]. A dynamic updating procedure, which averages the reference and test feature sets upon successful verification, is also essential for maintaining a high user acceptance rate over time. This is especially helpful during the first 10 to 15 sessions, when a person is learning to use the system efficiently.

An important application for speaker verification is in controlling access to automatic telephone transactions, such as banking and credit card transactions over the telephone, voice mail, etc. In this environment, the speech signal is distorted by the transmission channel and the variabilities in the telephone handset microphones. In addition, different types of digital coding of speech for transmission may degrade the performance [33]. Furthermore, lack of control over operational conditions such as user training and cooperation can adversely affect the performance of the system and its acceptance. Nevertheless, with the proliferation of business transactions over the phone using voice Input/Output (I/O), there is an increasing interest in deploying speaker verification systems over the public telephone network; and recent studies have demonstrated considerable progress [8] [24] [27]. Such a system can enhance security, save costs by reducing operator burden, and provide a personalized repertoire of voice services.

System Performance

The usefulness of a speaker verification system to a specific application is determined by the following factors:

- True-speaker rejection rate and impostor acceptance rate
- Verification throughput
- Enrollment
- Reference file storage
- System cost
- Mimic resistance
- Human factors design

The true-speaker rejection rate (type I error) and the impostor acceptance rate (type II error) are the primary parameters of system performance. For most entry control applications, a type I error rate of 1% and a type II error rate of 0.1% are acceptable. A typical set of type I and type II error curves as a function of acceptance threshold T is shown in Figure 3. The true-speaker and impostor data were obtained from nine speakers who used the password, "015678 Stockholm Canyon." There were a total of 540 true speaker (*tr sp*) trials and 648 impostor (*imp*) trials. The speech database was collected using a high-quality microphone in a quiet room. Error probability refers to the probability of false acceptance or false rejection. x is the verification score, which is compared against a decision threshold T for acceptance or rejection.

There is a tradeoff between these two parameters that is controlled by varying the decision thresholds. The operating values of the type I and type II error rates are determined by the particular application. In a high-security portal access control environment such as a military base or a computer center, which typically has 50 to 1,000 users, it is essential to maintain a very low impostor acceptance rate (less than 0.1%). The consequent higher true-speaker rejection rate (2%–3%) can be tolerated since the user population is small and the effort involved in handling these rejections is low. However, in telephone applications, where the user population is large

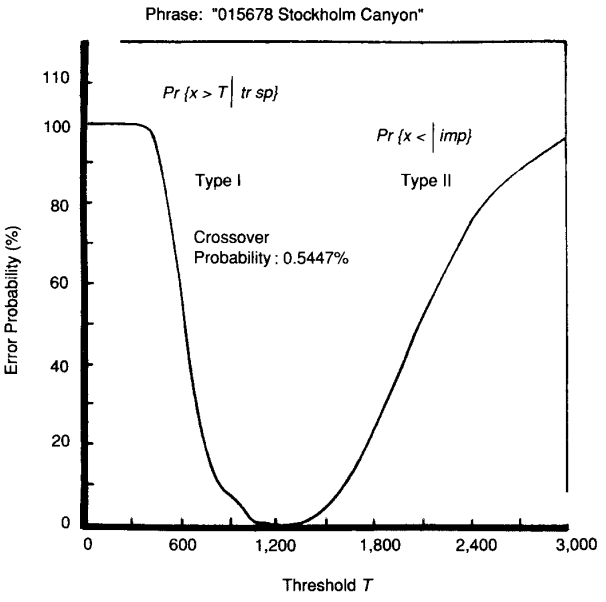


Fig. 3. Template-based, fixed-text speaker verification system error curves.

(thousands to millions) and little control exists over the user environment, the true-speaker rejection rate should be very low (less than 1%), since the cost of handling rejections is prohibitive and repeated rejections will alienate the user. The consequent higher impostor acceptance rate (5%–10%) is tolerable, since this level of impostor acceptance is still a vast improvement over conventional means of identity verification over the telephone, such as proffering, via telephone keypad, a personalized password made up of digits. It should also be noted that speaker verification restricts the access to a protected resource to one individual and thereby curtails the distribution of stolen passwords, or a similar abuse mechanism.

The duration of a transaction is determined by the number of repetitions of the verification phrase and its duration.

Verification throughput is determined by the duration of one verification transaction. The duration of a transaction is determined by the number of repetitions of the verification phrase and its duration. A typical transaction is between 5 and 20 s long, measured from the moment of identity claim (proffering a personal identification number) to the granting of access. While a high throughput rate is desirable, applications requiring high levels of security will need more speech data and a longer transaction time. The throughput time is also affected by whether or not a voice prompt is used. Prompts are necessary when a random set of phrases is used. Using the same phrase for each transaction will result in consistency and higher throughput but may make it easier for intruders to defeat the system using audio recordings.

Enrollment involves creation of the initial reference template for each user and subsequent updates of these templates. It usually lasts from 5 s to 10 minutes. Enrollment time should be kept low, particularly in telephone applications where hold times over the network are expensive. The reference file for each user contains the reference templates, update information, and administrative data such as times and dates of authorization, area of access, etc. The file size, in number of bytes, determines how many users can effectively be handled by the system. Typically, the size of a template is 1,000–5,000 bytes/user. With the availability of high-speed, low-cost digital signal processors, the cost of speaker verification systems is steadily decreasing. The current trend is towards providing integrated solutions, such as speech recognition, voice store and forward, and speech synthesis along with speaker verification on a common platform.

The resistance of operational systems to professional mimics has been studied in a few restricted scenarios. In one laboratory study [4], professional mimics were more successful than casual impostors. But the degree of success was also a function of the speech features used and improvements to the features resulted in a substantial reduction in the impostor acceptance rate for these mimics. Acceptance of a speaker verification system by a user population is largely determined by its user-friendliness. For example, what happens if the user is rejected? How are temporary vocal disorders such as laryngitis handled? Can the user obtain operator assistance if rejected by the system? What types of user instructions and prompts are employed? These human factor issues must be given careful consideration in the design of an operational system.

Databases

As speaker verification systems become operationally feasible, we see an acute need for standardized procedures for evaluating and comparing the performance of different systems. Comparison of fixed-text verification systems is made difficult by the fact that each system uses its own protocols, such as type of voice password, sequential strategy, enrollment and update criteria, etc. Free-text systems are not constrained by these parameters, but they are not well represented among the operational systems available today.

Several experimental databases have been used for the evaluation of different speaker verification systems [6] [8] [27] [34] [35], with specific protocols and experimental paradigms. But a set of benchmark databases would serve the purpose of objectively evaluating the merits of each system and providing new directions for further research and development. Such a database should include a large number of speakers, a variety of dialects, and voice passwords. It should represent the types of channel and handset variations encountered in practical systems. It should include changes in the speech of an individual speaker over time. Impostor testing of fixed-text systems requires that several users share voice passwords.

An Example of a Speaker Verification System

Texas Instruments has fielded several generations of speaker verification systems for access control applications. One such system has been in operation since 1974, controlling access to their computer center premises. It uses an algorithm based on a 14-channel filter-bank representation for the speech signal and a Euclidian metric between the reference and input speech tokens, using a dynamic time-warping alignment procedure [2] [15]. The verification protocol consists of a four-word fixed phrase and each of these four words is randomly chosen from a set of four possible words in each word position. A multiple phrase strategy is used, which allows up to seven phrases, but in operational use the average number of phrases was 1.6. The gross rejection rate on this system has been measured at 0.9% (this includes rejection due to all sources) for an operational system with a casual impostor acceptance rate of 0.7%.

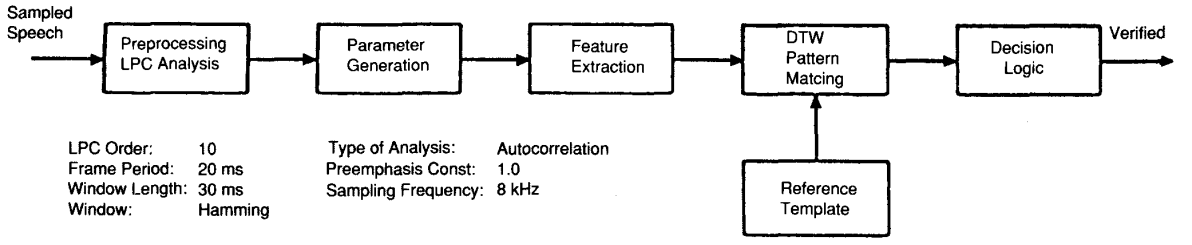


Fig. 4. Verification algorithm.

Texas Instruments has recently developed a fixed-text speaker verification system for access control that can be implemented on a small portable professional computer. The verification protocol consists of a primary phrase made up of five digits followed by a two-word coda; for example, "32109 Jackson Basin." The use of a fixed text phrase reduces user learning problems and assures greater consistency in speaking. The length of the speech material can be conveniently increased, yielding better discriminability between valid users and impostors. A secondary phrase, consisting of a random sequence of five digits, is also used to prevent the use of tape recorders by intruders. Successful verification is contingent upon acceptance of the user at both the primary and secondary verification stages.

The linear predictive analysis front-end is used for modeling the digitized speech data, and a simulated filter-bank on a critical band frequency scale is derived from the LPC spectrum. The filter-bank amplitudes, in dB, are then transformed to a set of uncorrelated features using a statistically optimum linear transformation [9]. The resulting Principal Spectral Components (PSCs) are ranked in decreasing order of their statistical variance and the least significant features are discarded. In this system, a 14-element filter-bank magnitude vector is transformed to a 10-element PSC vector. The Euclidean distance is used to compare the test and reference sets of PSC vectors using a dynamic time-warping algorithm. The resulting distance, averaged over the duration of the utterance, is compared against a decision threshold. The block diagram of the algorithm is in Figure 4. The digits 0–9 for the secondary phrase are enrolled by automatically segmenting each individual digit from a five-digit sequence, using speaker independent digit templates. Each verification session consists of up to three utterances of the primary phrase and up to two utterances of the secondary phrase.

A large-scale evaluation of this system was performed over a four-month period. The system used for performance evaluation consisted of a host, which performed control functions, report generation, and error analysis; and a terminal, which performed all the speech processing functions. The reference templates were stored on the host and downloaded to the terminal on request. The speech processing functions were performed on a speech subsystem comprising a TMS32010-based Texas Instruments (TI) Speech Command System and a custom-built analog I/O board. A TI Professional Computer (TIPC) was used as a host and two portable TIPCs were used as terminals for simultaneous use. A population of 100 men and 100 women of various dialectal and educational backgrounds participated in the evaluation over a four-month period. Each user contributed 40 sessions, with two true speaker trials and six impostor trials per session, both performed on-line. Both visual and aural prompts were used. Every user attempted impostor trials against every other user at least once. No cross-gender impostor trials were allowed. A reenrollment strategy was employed if a user was rejected several times consecutively according to a predefined criterion. The decision thresholds were derived from a pilot database of forty speakers, and only one of them was present in both training and test populations. A speaker normalization technique was used, which normalized the decision thresholds based on the prior performance of each user and a global average score. This increased the true-speaker acceptance rate, without noticeably increasing the impostor acceptance rate. The reference template for the true speaker was linearly averaged with the test utterance only upon successful verification. The improvement in performance over time is illustrated in Figure 5. The overall system performance for post enrollment sessions (after each user has had four successful acceptances and is comfortable with system use) is shown in Table I.

Each user was allowed up to three utterances of the primary phrase and two utterances of the secondary phrase. The percentage acceptance as a function of the number of utterances of the primary phrase is summarized in Table II.

A small percentage of users will always have difficulty with the system. (These problem users are sometimes referred to as "goats," while those with consistently good performance qualify to be labelled as "sheep"!) These users must be re-enrolled several times. In the above evaluation, less than 10% of the user population had four or more rejections (were "goats") during the forty-session trial, as shown in Figure 6.

Industry Activity

Several other research organizations have evaluated their own verification systems. AT&T Bell Laboratories have designed and evaluated several generations of speaker verification systems [4]. Recently, they have tested a system for use

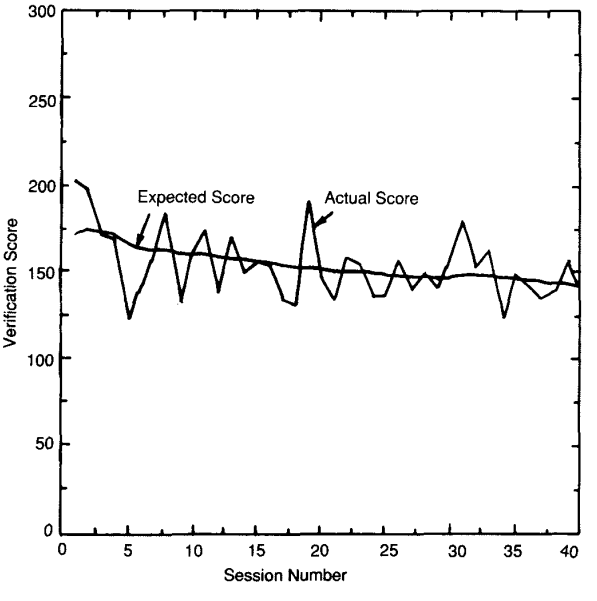


Fig. 5. Performance improvement over time.

Table I. Overall System Performance*

	MALE	FEMALE
True-Speaker Rejection	54/6,760 (0.80%)	51/5,829 (0.87%)
Impostor Acceptance	15/20,949 (0.07%)	23/19,400 (0.12%)

*For post-enrollment sessions

Table II. Primary-Phase Acceptance

Number of Phrases	Acceptance (%)	
	Men	Women
1	92.5	89.3
2	5.6	7.8
3	1.1	1.7

over the telephone lines [8]. This system resulted in an equal error rate (the point at which type I and type II error rates are equal for a given set of decision thresholds) of 1.9%. The distortions introduced by differences in the microphones (carbon, Electret, and dynamic types) used in telephone handsets are a major hurdle in this application. Among the organizations that have research and development programs in speaker verification systems are ITT Defense Communications Division, Bell Communications Research, Siemens Research and Technology Laboratories, and the Regional Bell Operating Companies [21–23].

There has been a growing interest in the commercial applications of speaker verification as seen from product offerings. Currently, several companies (Voxtron, Inc., ECCO Industries, and Alpha Microsystems, to name a few) have marketed speaker verification products. These include systems for use over the telephone lines as well as for portal access control [23] [24]. Sandia National Laboratories conduct evaluations of different types of personal identity verifiers, including speaker verification systems. A recent study [25] evaluated two speaker verification systems: the AT&T Voice Password Verifier System, supplied by AT&T Information Systems and built specifically for the evaluation (not a commercial product); and the Veritron 1000, a commercial product marketed by Voxtron. The AT&T system, which was tested over the public phone lines, resulted in a false rejection rate of 3.3% after the second try (a maximum of two tries per session allowed) and a false acceptance rate of 0.42% after two tries. The average verification transaction time was 8.8 s measured from "Off-hook" to "Access granted." The enrollment time was 0.3 minutes. The Voxtron system used a TI-Business Pro professional computer and a standard telephone handset. The false rejection rate was 5.2% after the third try (a maximum of three tries allowed) and a false acceptance rate of 0.91% after the third try, when tested over the public phone lines. The average verification time was 10.1 s and the enrollment time was 2.4 minutes for enrolling four phrases for each user.

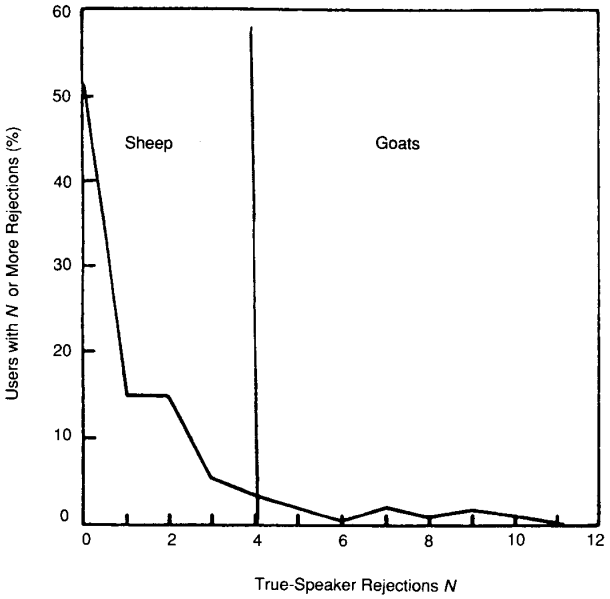


Fig. 6. "Goat" and "sheep" distribution.

Summary

In the last decade, speaker verification systems have achieved significant improvements in performance, cost effectiveness, and user acceptability. Fixed-text speaker verification systems have been introduced into the market and are finding many applications in personal identity verification. It is encouraging to see active interest in telephone applications, since this is where speaker verification technology can have a significant impact. As the technology matures, we need to define utilitarian measures of system performance and objective methods of system evaluation.

Fixed-text speaker verification systems have been introduced into the market and are finding many applications in personal identity verification.

Acknowledgments

The author gratefully acknowledges the help provided by the Speech and Image Understanding Laboratory at Texas Instruments during the preparation of this tutorial. In particular, the technical guidance of Dr. George R. Doddington is sincerely appreciated.

References

[1] B. S. Atal, "Automatic Recognition of Speakers from their Voices," *Proc. IEEE*, vol. 64, no. 4, pp. 460–475, Apr. 1976.

[2] G. R. Doddington, "Speaker Recognition—Identifying People by their Voices," *Proc. IEEE*, vol. 73, pp. 1,651–1,664, Nov. 1985.

[3] M. R. Sambur, "Selection of Acoustic Features for Speaker Identification," *IEEE Trans. Acoust., Speech, Sig. Processing*, vol. ASSP-23, pp.176–182, Apr. 1975.

[4] A. E. Rosenberg, "Automatic Speaker Verification: A Review," *Proc. IEEE*, vol. 64, pp. 475–487, Apr. 1976.

[5] B. S. Atal, "Automatic Speaker Recognition Based on Pitch Contours," *J. Acoust. Soc. Amer.*, vol. 52, pp. 1,687, Dec. 1972.

[6] A. H. Gray and J. D. Markel, "Distance Measures for Signal Processing," *IEEE Trans. Acoust., Speech, Sig. Processing*, vol. ASSP-24 pp. 380–391, 1975.

[7] J. M. Naik and G. R. Doddington, "High Performance Speaker Verification Using Principal Spectral Components," *Proc. ICASSP '86*, pp. 881–884, 1986.

[8] M. R. Birnbaum, L. A. Cohen, and F. X. Welsh, "A Voice Password System for Access Security," *AT&T Tech. J.*, vol. 65, no. 5, pp. 68–74, Sept./Oct. 1986.

[9] P. K. Rajasekaran and G. R. Doddington, "Speech Recognition in the F-16 Cockpit Using Principal Spectral Components," *Proc. ICASSP '85*, pp.882–885, 1985.

[10] G. R. Doddington, "A Method of Speaker Verification," *J. Acoust. Soc. Amer.*, vol. 49, pt. 1, p. 139(A), Jan. 1971.
[11] R. C. Lummis, "Speaker Verification by Computer Using Speech Intensity for Temporal Registration," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 80–89, 1973.
[12] S.S. McCandless, "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra," *IEEE Trans. Acoust., Speech, Sig. Processing*, vol. ASSP-22, pp. 135–141, Apr. 1974.
[13] P. K. Rajasekaran, "Real-Time Factoring of the Linear Prediction Polynomial of Speech Signals," *Dig. Sig. Processing '84*, V. Cappallini and A. G. Constantinidis, eds., Elsevier Science Publishers, pp. 405–410, 1984.
[14] L.-S. Su, K. P. Li, and K. S. Fu, "Identification of Speakers by Nasal Coarticulation," *J. Acoust. Soc. Amer.*, vol. 156, pp. 1,876–1,882, Dec. 1974.
[15] G. R. Doddington, "Speaker Verification," Final Rep. RADC-TR-74-179, Griffis Air Force Base, Rome, NY, 1974.
[16] S. K. Das and W. S. Mohn, "A Scheme for Speech Processing in Automatic Speaker Verification," *IEEE Trans. Audio Electroacoust.*, vol. AU-19, pp. 32–43, Mar. 1971.
[17] M. R. Sambur, "Speaker Recognition Using Orthogonal Linear Prediction," *IEEE Trans. Acoust., Speech, Sig. Processing*, vol. ASSP-24, pp. 283–289, Mar. 1976.
[18] R.E. Wohlford, "A Comparison of Four Techniques for Automatic Speaker Recognition," *Proc. ICASSP '80*, pp. 908–911, 1980.
[19] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," *IEEE Trans. Acoust., Speech, Sig. Processing*, vol. ASSP-29, no. 2, pp. 254–272, Apr. 1981.
[20] M. DeGeorge and W. Feix, "A Speaker Verification for Access Control," *Proc. of Speech Tech '86*, pp. 281–286, Apr. 1986.
[21] A. L. Higgings and R. E. Wohlford, "A New Method of Speaker Recognition," *Proc. ICASSP '86*, Apr. 1986.
[22] E. H. Wrench, "A Real-Time Implementation of a Text Independent Speaker Recognition System," *Proc. ICASSP '81*, pp. 193–196, Apr. 1981.
[23] S. Chapin, Voxtron, Inc., New Braundfels, TX, personal communication.
[24] K. Paley, Ecco Industries, Inc., Danvers, MA, personal communication.
[25] R. L. Maxwell and L. J. Wright, "A Performance Evaluation of Personnel Identity Verifiers," *Proc., Ann. Mtg. of the Inst. of Nuclear Materials Mgmt.*, July 1987.
[26] W. D. Voiers, "Perceptual Bases of Speaker Identity," *J. Acoust. Soc. Amer.*, vol. 36, no. 6, pp. 1,065–1,073, 1964.
[27] J. M. Naik, L. P. Netsch, and G. R. Doddington, "Speaker Verification over Long Distance Telephone Lines," *Proc. ICASSP '89*, pp. 524–527, May 1989.
[28] F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Trans. Acoust., Speech, Sig. Processing*, vol. ASSP-23, pp. 67–72, 1975.
[29] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE Trans. Acoust. Speech, Sig. Processing*, vol. ASSP-26, pp. 43–49, 1978.
[30] D. O'Shaughnessy, "Speaker Recognition," *IEEE ASSP Mag.*, pp. 4–17, Oct. 1986.
[31] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[32] J. Picone, G. R. Doddington, and J. J. Godfrey, " A Layered Grammar Approach to Speaker Independent Speech Recognition," *1988 IEEE Speech Recog. Workshop*, Harriman, NY, June 1988.
[33] C. McGonegal, A. Rosenberg, and L. R. Rabiner, "The Effects of Several Transmission Systems on Automatic Speaker Verification Systems," *Bell Syst. Tech. J.*, vol. 58, pp. 2,071–2,087, 1979.
[34] J. M. Naik and G. R. Doddington, "Evaluation of a High Performance Speaker Verification System for Access Control," *Proc. ICASSP '87*, pp. 2,392–2,395, Apr. 1987.
[35] G. Velius, "Variants of Cepstral Based Speaker Identity Verification System," *Proc. ICASSP '88*, pp. 583–586, Apr. 1988.

Biography

Jayant M. Naik received the Ph.D. in electrical engineering from the University of Florida, Gainesville, FL, the M.S. in electrical engineering from Wichita State University, Kansas, and the B.E. in electronics engineering from Bangalore University, Bangalore, India.

He has been working in the area of speaker verification for the past six years, from 1984 to 1989, previously at the Speech and Image Understanding Laboratories of Texas Instruments, Dallas, Texas, and presently at the NYNEX Science and Technology Center, White Plains, NY. He has published over twenty papers in the areas of speaker verification, speech analysis/synthesis, and speech production modeling, and has two patents in speaker verification pending.

Dr. Naik was a Chairman of the Dallas chapter of the IEEE ASSP Society from 1985 to 1987, and is a member of Tau Beta Pi and Eta Kappa Nu.