Automatic Speaker Recognition Using Gaussian Mixture Speaker Models

Douglas A. Reynolds

■ **Speech conveys several levels of information. On a primary level, speech conveys the words or message being spoken, but on a secondary level, speech also reveals information about the speaker. The Speech Systems Technology group at Lincoln Laboratory has developed and experimented with approaches for automatically recognizing the words being spoken, the language being spoken, and the topic of a conversation. In this article we present an overview of our research efforts in a fourth area—automatic speaker recognition. We base our approach on a statistical speaker-modeling technique that represents the underlying characteristic sounds of a person's voice. Using these models, we build speaker recognizers that are computationally inexpensive and capable of recognizing a speaker regardless of what is being said. Performance of the systems is evaluated for a wide range of speech quality, from clean speech to telephone speech, by using several standard speech corpora.**

TASKS THAT ARE EASILY PERFORMED by humans, such as face or speech recognition, prove difficult to emulate with computers. Speaker-recognition technology stands out as one application in which the computer outperforms the humans.

For over six decades, scientists have studied the ability of human listeners to recognize and discriminate voices [1]. By establishing the factors that convey speaker-dependent information, researchers have been able to improve the naturalness of synthetic and vocoded speech [2] and assess the reliability of speaker recognition for forensic science applications [3]. Soon after the development of digital computers, research on speaker recognition turned to developing objective techniques for automatic speaker recognition, which quickly led to the discovery that simple automatic systems could outperform human listeners on a similar task [4].

Over the last three decades, researchers have developed increasingly sophisticated automatic speaker-recognition algorithms, and the performance of these algorithms in more realistic evaluation speech corpora has improved. Today, task-specific speaker-recognition systems are being deployed in large telecommunications applications. For example, in 1993 the Sprint Corporation offered the Voice FoneCard calling card, which uses speaker recognition to allow access to its long-distance network.

The general task of automatic speaker recognition is far from solved, however, and many challenging problems and limitations remain to be overcome. In this article we present an overview of the research, developments, and evaluation of automatic speaker-recognition systems at Lincoln Laboratory.

Problem Definition and Applications

Speaker recognition involves two tasks: identification and verification, as shown in Figure 1. In identification, the goal is to determine which voice in a known group of voices best matches the speaker. In verification, the goal is to determine if the speaker is who he or she claims to be.

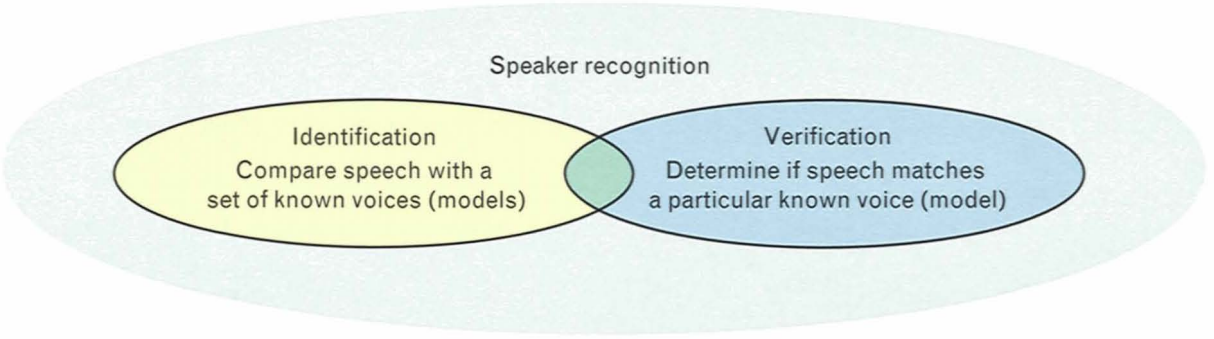


FIGURE 1. The two fundamental tasks of speaker recognition: identification and verification. The goal of speaker recognition is to recognize a person automatically from his or her voice. In identification, the incoming speech is compared with a set of known voices. In verification, the incoming speech is compared with one known voice.

In speaker identification, the unknown voice is assumed to be from the predefined set of known speakers. For this type of classification problem—an *N*-alternative, forced-choice task—errors are defined as misrecognitions (i.e., the system identifies one speaker's speech as coming from another speaker) and the difficulty of identification generally increases as the speaker set (or speaker population) increases.

Applications of pure identification are generally unlikely in real situations because they involve only speakers known to the system, called enrolled speakers. However, one indirect application of identification is speaker-adaptive speech recognition, in which speech from an unknown speaker is matched to the most similar-sounding speaker already trained on the speech recognizer [5]. Other potential identification applications include intelligent answering machines with personalized caller greetings [6] and automatic speaker labeling of recorded meetings for speaker-dependent audio indexing [7, 8].

Speaker verification requires distinguishing a speaker's voice known to the system from a potentially large group of voices unknown to the system. Speakers known to the system who claim their true identity are called *claimants*; speakers, either known or unknown to the system, who pose as other speakers are called *impostors*. There are two types of verification errors: false acceptances—the system accepts an impostor as a claimant; and false rejections—the system rejects a claimant as an impostor.

Verification forms the basis for most speaker-recognition applications. Current applications such as computer log-in, telephone banking [9, 10], calling cards, and cellular-telephone fraud prevention substitute or supplement a memorized personal identification code with speaker verification. Verification can also be applied as an information retrieval tool for retrieving messages from a voice mailbox.

Speaker-recognition tasks are further distinguished by the constraints placed on the text of the speech used in the system [3]. In a *text-dependent system*, the spoken text used to train and test the system is constrained to be the same word or phrase. For example, in an access-control verification application a claimant can always use the same personalized code. Thus a speaker-verification system can take advantage of knowing the text to be spoken. Such a verification system can be fooled, however, by recording a claimant's phrase and playing it back to gain access. In a *text-independent system*, training and testing speech is completely unconstrained. This type of system is the most flexible and is required for applications such as voice mail retrieval, which lacks control over what a person says.

Between the extremes of text dependence and text independence falls the *vocabulary-dependent system*, which constrains the speech to come from a limited vocabulary, such as the digits (e.g., "zero," "one") from which test words or phrases (e.g., "zero-one-eight") are selected. This system provides more flexibility than the text-dependent system because pass phrases used by claimants can be changed regularly without retraining to help thwart an impostor with a tape recorder.

Features for Speaker-Recognition Systems

To develop machines for speaker recognition, scientists and engineers must first ask, "How do humans recognize one another by voice alone?" We use many perceptual cues, some nonverbal, when recognizing speakers. These cues are not well understood, but range from high-level cues, which are related to semantic or linguistic aspects of speech, to low-level cues, which are related to acoustic aspects of speech.

High-level cues include word usage, idiosyncrasies in pronunciation, and other nonacoustic characteristics that can be attributed to a particular speaker. These cues describe a person's manner of speech and are generally thought to arise from varied life experiences, such as place of birth and level of education. These cues are also termed learned traits. Low-level cues, on the other hand, are more directly related to the sound of a person's voice and include attributes such as soft or loud, clear or rough, and slow or fast.

While human listeners use all levels of cues to recognize speakers, low-level cues have been found to be the most effective for automatic speaker-recognition systems. Low-level cues can be related to acoustic measurements that are easily extracted from the speech signal. On the other hand, high-level cues are not easily quantified, and can occur infrequently in text-independent speech and not at all in text-dependent speech. They are also difficult to extract from the speech signal—looking for certain words would require a reliable speech recognizer or word spotter.

To find acoustic measurements from a speech signal that relate to physiological attributes of the speaker, we consider the basic model of speech production. In this model, speech sounds are the product of an air stream passed through the glottis, producing resonances in the vocal tract and nasal cavities. During voiced sounds, such as vowels, the glottis rhythmically opens and closes to produce a pulsed excitation to the vocal tract. During unvoiced sounds, such as fricatives, the glottis remains partially open, creating a turbulent airflow excitation. To produce different sounds, the vocal tract moves into different configurations that change its resonance structure. Nasal sounds are produced by shunting the glottal excitation through the nasal cavities.

From this model we see that the glottis and vocal tract impart the primary speaker-dependent characteristics found in the speech signal. The periodicity, or pitch, of the speech signal contains information about the glottis. Major frequency components of the speech spectrum contain information about the vocal tract and nasal cavities. Speech spectral information from the frequency components has proven to be the most effective cue for automatic speaker-recognition features. Although pitch conveys speaker-specific information and can be used in some controlled applications, it can be difficult to extract reliably, especially from noise-corrupted speech, and it is more susceptible to nonphysiological factors such as the speaker's emotional state and level of speech effort.

Figure 2 shows examples of how vocal-tract configurations produce different spectra for two steady-state vowel sounds. The top part of the figure shows the cross section of the vocal tract. Below is a plot of the frequency spectrum (magnitude versus frequency) for the vowel sound. The peaks in the spectrum are resonances produced by the particular vocal-tract configuration and are known as the speech formants. For each vocal-tract configuration, we show the spectrum for two different speakers: a male and a female.

Note that for any particular sound, the relative location of the formants within each speaker's spectrum is similar, since the same sound is being produced. By comparing the speaker's spectra, however, we see that corresponding formants occur at different frequencies and with different intensities—a direct result of the different vocal-tract structures. Most automatic speaker-recognition systems rely upon these spectral differences to discriminate speakers.

Natural speech is not simply a concatenation of sounds. Instead, it is a blending of different sounds, often with no distinct boundaries between transitions. Figure 3 shows the digitally sampled speech waveform of a continuously spoken sentence and the corresponding spectra. The spectra are presented as a three-dimensional time-frequency spectrogram with frequency on the y-axis, time on the x-axis, and darker regions representing higher spectral energy. The spectrogram illustrates the dynamic nature of the formants (seen as dark bands in the spectrogram) and hence the vocal tract.

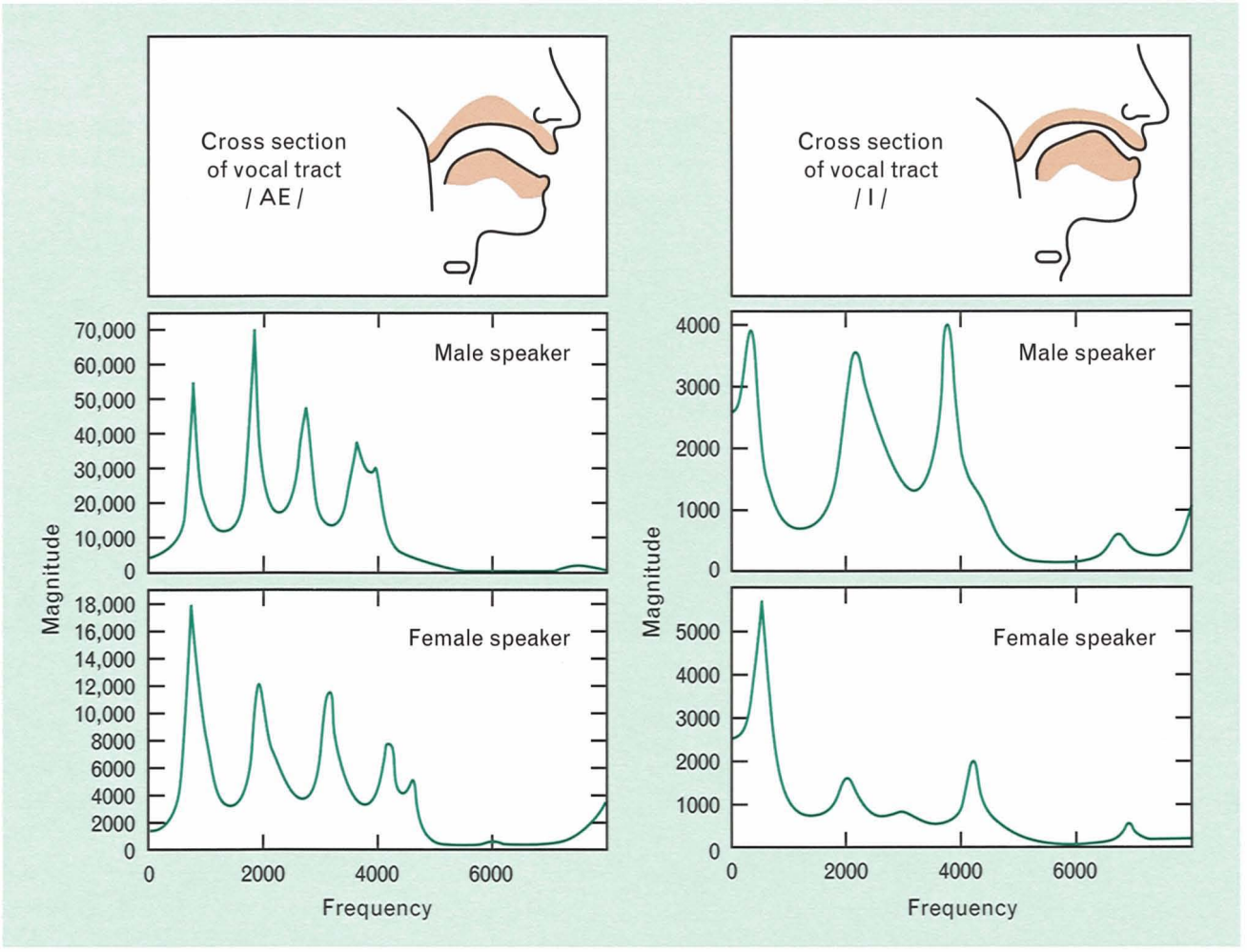


FIGURE 2. Examples of vocal-tract configurations and the corresponding frequency spectra from two steady-state vowels spoken by two different speakers: a male and a female. The peaks, or formants, in the spectra are resonances produced by the particular vocal-tract configuration.

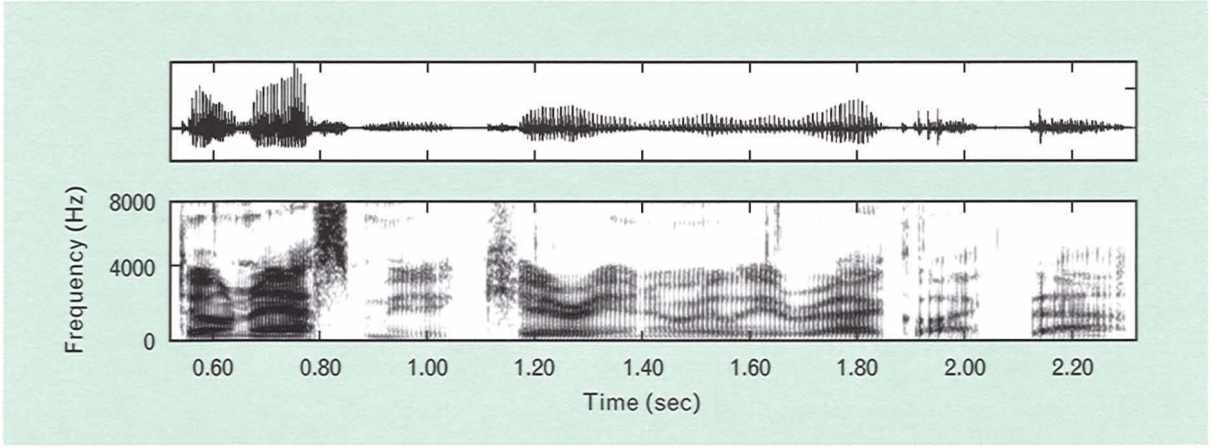


FIGURE 3. Digitally sampled speech waveform of a spoken sentence (above) and corresponding spectrogram (below) showing the dynamic nature of the formants as the vocal tract continuously changes shape. The sentence spoken was "Don't ask me to carry an oily rag like that."

To obtain steady-state measurements of the spectra from continuous speech, we perform short-time spectral analysis, which involves several processing steps, as shown in Figure 4. First, the speech is segmented into frames by a 20-msec window progressing at a 10-msec frame rate. A speech activity detector is then used to discard silence and noise frames [11, 12]. For text-independent speaker recognition, removing silence and noise frames from the training and testing signals is important in order to avoid modeling and detecting the environment rather than the speaker.

Next, spectral features are extracted from the speech frames. A reduced spectral representation is produced by passing the speech frame through a pseudo filter bank designed to match the frequency sensitivity of the ear. This type of filter bank is called a mel-scale filter bank and is used extensively for speech-recognition tasks [13]. Passing the speech frame through a pseudo filter produces a spectral representation consisting of log magnitude values from the speech spectrum sampled at a linear 100-Hz spacing below 1000 Hz and sampled at a logarithmic spacing above 1000 Hz.

For 4-kHz bandwidth speech (e.g., telephone-quality speech), this reduced spectral representation has twenty-four log magnitude spectrum samples. The log magnitude spectral representation is then inverse Fourier transformed to produce the final representation, called cepstral coefficients. The last transform is used to decorrelate the log magnitude spectrum samples. We base the decision to use mel-scale cepstral coefficients on good performance in other speech-recognition tasks and a study that compares several standard spectral features for speaker identification [14].

The sequence of spectral feature vectors extracted from the speech signal is denoted $\{\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T\}$, where the set of cepstral coefficients extracted from a speech frame are collectively represented as a D-dimensional feature vector $\mathbf{x}_t$, and where t is the sequence index and T is the number of feature vectors.

Finally, the spectral feature vectors undergo channel compensation to remove the effects of transmission degradation. Caused by noise and spectral distortion, this degradation is introduced when speech travels through communication channels like telephone or cellular phone networks.

The resulting spectral sequence representation is the starting point for almost all speech-related tasks, including speech recognition [15] and language identification [16]. Unfortunately, this representation is not a particularly efficient representation for speaker recognition. Much of a spectral sequence represents the linguistic content of the speech, which contains large redundancies and is mostly not needed for speaker representation.

Statistical Speaker Model

Specific speaker-recognition tasks are accomplished by employing models that extract and represent the desired information from the spectral sequence. Since the primary speaker-dependent information conveyed by the spectrum is about vocal-tract shapes, we wish to use a speaker model that in some sense captures the characteristic vocal-tract shapes of a person's voice as manifested in the spectral features. Because of

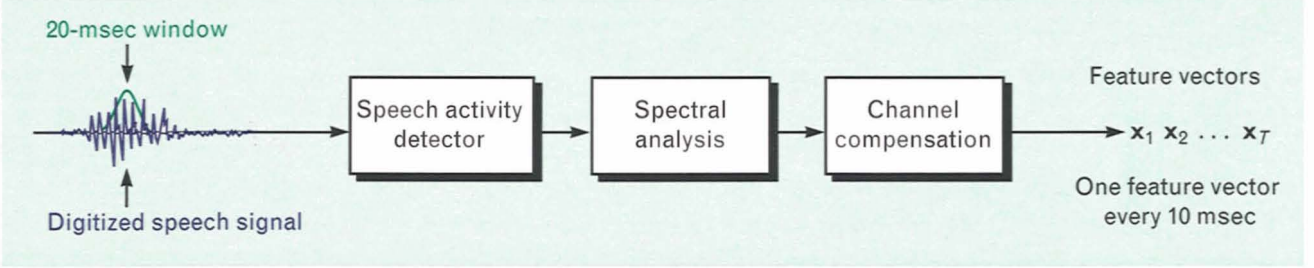


FIGURE 4. Front-end signal processing used to produce feature vectors from the speech signal. Twenty-msec segments, or frames, of speech are passed through a speech activity detector, which discards silence and noise frames that reflect the environment rather than the speaker. Spectral analysis extracts spectral features from the speech frames. Channel compensation removes the effects of transmission degradation from the resulting spectral representations.

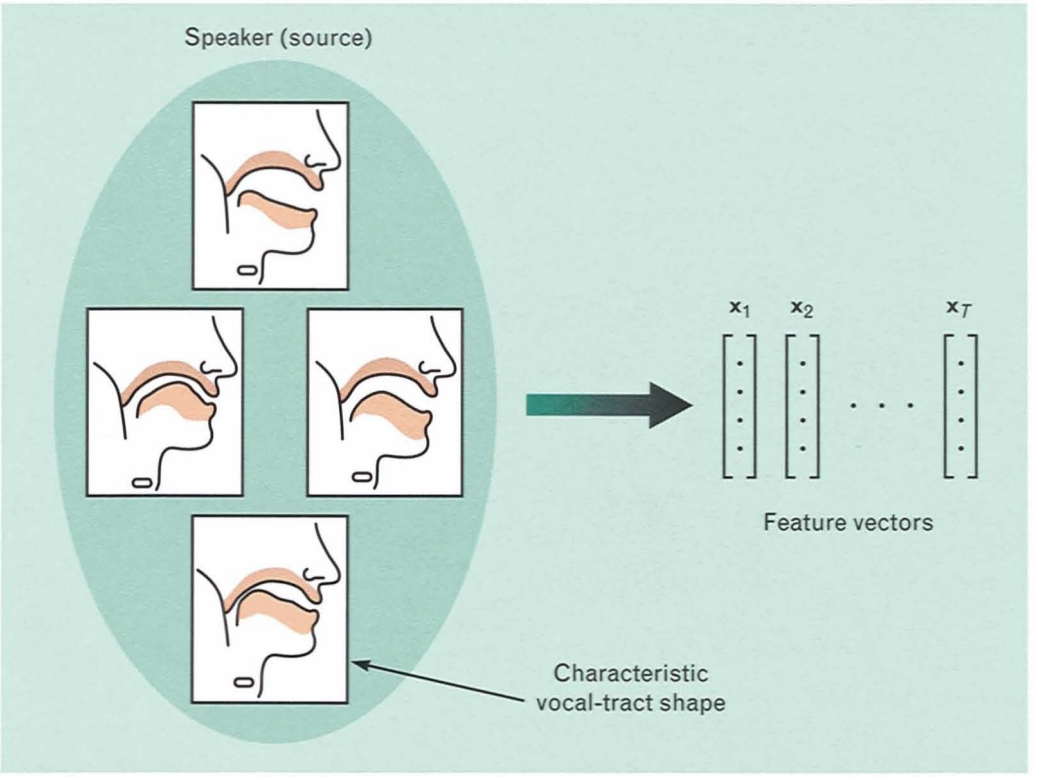


FIGURE 5. Statistical speaker model. The speaker is modeled as a random source producing the observed feature vectors. Within the random source are states corresponding to characteristic vocal-tract shapes.

the success of statistical pattern-recognition approaches for a wide variety of speech tasks, we adapt a statistical formulation for such a speaker model.

In the statistical speaker model, we treat the speaker as a random source producing the observed feature vectors, as depicted in Figure 5. Within the random speaker source, there are hidden states corresponding to characteristic vocal-tract configurations. When the random source is in a particular state, it produces spectral feature vectors from that particular vocal-tract configuration. The states are called hidden because we can observe only the spectral feature vectors produced, not the underlying states that produced them.

Because speech production is not deterministic (a sound produced twice is never exactly the same) and spectra produced from a particular vocal-tract shape can vary widely due to coarticulation effects, each state generates spectral feature vectors according to a multidimensional Gaussian probability density function (pdf), with a state-dependent mean and covariance. The pdf for state i as a function of the D-dimensional feature vector $\mathbf{x}$ is

$$b_i(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_i\right|^{1/2}} \times \exp\left\{-\frac{1}{2}\left(\mathbf{x}-\mu_i\right)^T\left(\Sigma_i\right)^{-1}\left(\mathbf{x}-\mu_i\right)\right\},$$

where μ_i is the state mean vector and Σ_i is the state covariance matrix. The mean vector represents the expected spectral feature vector from the state, and the covariance matrix represents the correlations and variability of spectral features within the state.

In addition to the feature-vector production being a state-dependent random source, the process governing what state the speaker model occupies at any time is modeled as a random process. The following discrete pdf associated with the M states describes the probability of being in any state,

$$\{p_1, \ldots, p_M\}, \text{ where } \sum_{i=1}^{M} p_i = 1,$$

and a discrete pdf describes the probability that a transition will occur from one state to any other state,

$$a_{ij} = \Pr(i \rightarrow j), \text{ for } i, j = 1, \ldots, M.$$

The above definition of the statistical speaker model is known more formally as an ergodic hidden Markov model (HMM) [17]. HMMs have a rich theoretical foundation and have been extensively applied to a wide variety of statistical pattern-recognition tasks in speech processing and elsewhere. The main motivation for using HMMs in speech-recognition tasks is that they provide a structured, flexible, computationally tractable model describing a complex statistical process.

Because we are primarily concerned with text-independent speech, we simplify the statistical speaker model by fixing the transition probabilities to be the same, so that all state transitions are equally likely. That is, we set a_{ij} equal to $1/M$. While the sequencing information of the states can contain some speaker-specific information, it generally represents linguistic information and has been shown experimentally to be unnecessary for text-independent speech [18].

The Gaussian Mixture Speaker Model

From the above definition of the statistical speaker model, we can show that the pdf of the observed spectral features generated from a statistical speaker model is a Gaussian mixture model (GMM) [19]. In terms of the parameters of an M-state statistical speaker model, the GMM pdf is

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\mathbf{x}), \qquad (1)$$

where

$$\lambda = (p_i, \mu_i, \Sigma_i), \text{ for } i = 1, \ldots, M$$

represents the parameters of the speaker model. Thus the probability of observing a feature vector $\mathbf{x}_t$ coming from a speaker model with parameter λ is the sum of the probabilities that $\mathbf{x}_t$ was generated from each hidden state, weighted by the probability of being in each state. With this summed probability we can produce a quantitative value, or score, for the likelihood that an unknown feature vector was generated by a particular GMM speaker model.

Despite the apparent complexity of the GMM, model parameter estimates are obtained in an unsupervised manner by using the expectation-maximization (EM) algorithm [20]. Given feature vectors extracted from training speech from a speaker, the EM algorithm iteratively refines model parameter estimates to maximize the likelihood that the model matches the distribution of the training data. This training does not require additional information, such as transcription of the speech, and the parameters converge to a final solution in a few iterations.

Applying the Model

With the GMM as the basic speaker representation, we can then apply this model to specific speaker-recognition tasks of identification and verification. The identification system is a straightforward maximum-likelihood classifier. For a reference group of S speaker models $\{\lambda_1, \lambda_2, \ldots, \lambda_S\}$, the objective is to find the speaker identity $\hat{s}$ whose model has the maximum posterior probability for the input feature-vector sequence $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$. The minimum-error Bayes' rule for this problem is

$$\hat{s} = \arg \max_{1 \le s \le S} \Pr(\lambda_s | X) = \arg \max_{1 \le s \le S} \frac{p(X|\lambda_s)}{p(X)} \Pr(\lambda_s).$$

Assuming equal prior probabilities of speakers, the terms $\Pr(\lambda_s)$ and $p(X)$ are constant for all speakers and can be ignored in the maximum. By using logarithms and assuming independence between observations, the decision rule for the speaker identity becomes

$$\hat{s} = \arg \max_{1 \le s \le S} \sum_{t=1}^{T} \log p(\mathbf{x}_t | \lambda_s),$$

in which T is the number of feature vectors and $p(\mathbf{x}_t|\lambda_s)$ is given in Equation 1. Figure 6(a) shows a diagram of the speaker-identification system.

Although the verification task requires only a binary decision, it is more difficult to perform than the

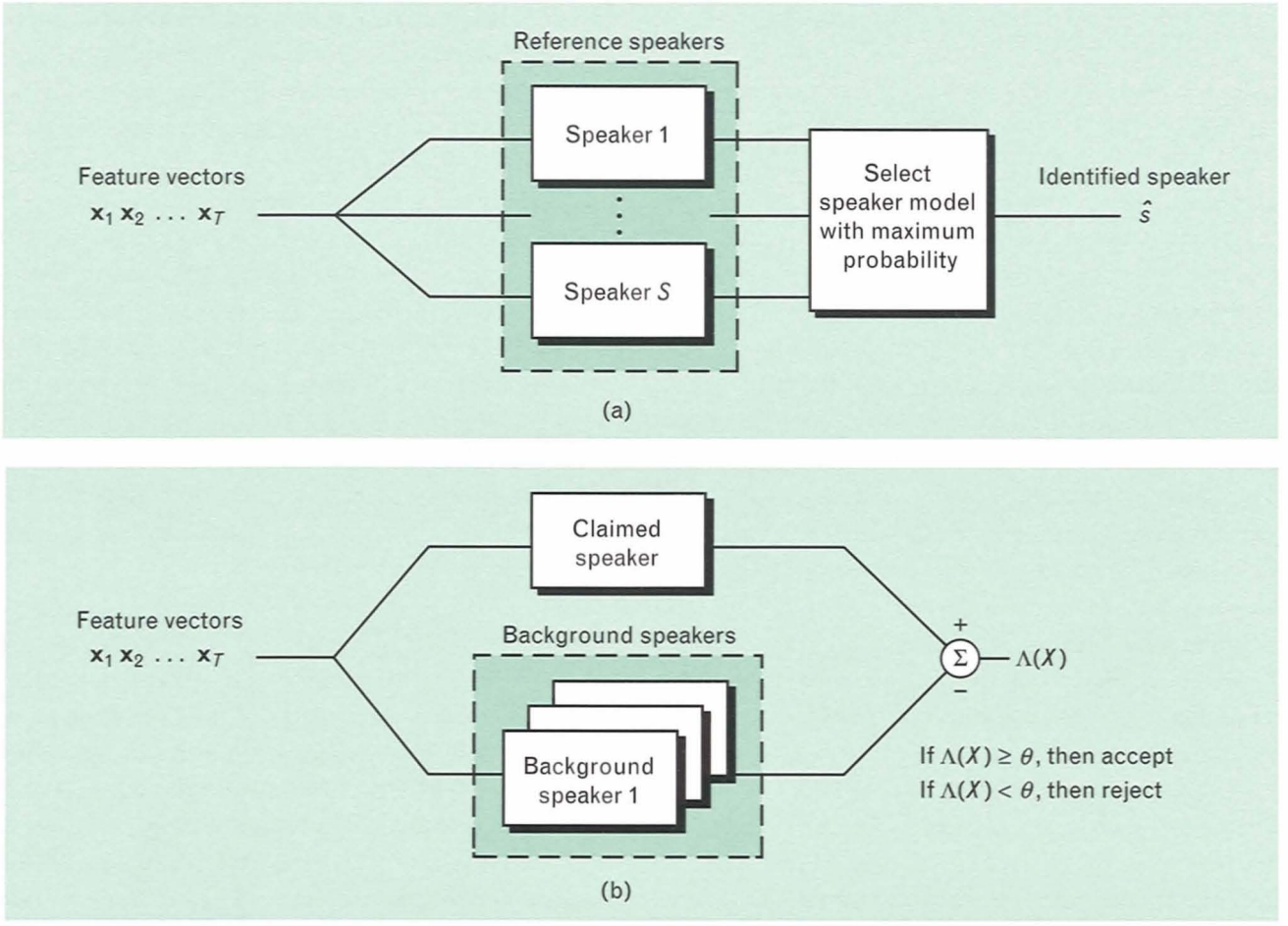


FIGURE 6. Speaker-recognition systems for identification and verification. The feature vectors extracted from the front-end processing in Figure 4 are fed into speaker identification and verification systems. (a) In identification, the goal is to pick the one speaker out of a group of S speakers whose model has the highest probability of generating the input feature vectors. (b) In verification, the system must decide if the input voice belongs to the claimed speaker or to another background speaker. The likelihood ratio $\Lambda(X)$ compares the score from the claimant model with that of a background model. Then the likelihood ratio is compared with a threshold value θ. The claimed speaker is accepted if $\Lambda(X) \geq \theta$ and rejected if $\Lambda(X) < \theta$.

identification task because the alternatives are less defined. Figure 6(b) shows a diagram of the speaker-verification system. The system must decide if the input voice belongs to the claimed speaker, having a well-defined model, or to some other speaker, having an ill-defined model. In a hypothesis-testing framework, for a given input utterance and a claimed identity, the choice becomes H_0 if X is from the claimed speaker, or H_1 if X is not from the claimed speaker.

A model of the universe of possible nonclaimant speakers must be used to perform the optimum likelihood-ratio test that decides between H_0 and H_1. The general approach used in the speaker-verification system is to apply a likelihood-ratio test to an input utterance to determine if the claimed speaker is accepted or rejected. For an utterance X, a claimed speaker identity with corresponding model λ_C, and the model of possible nonclaimant speakers $\lambda_{\bar{C}}$, the likelihood ratio is

$$\frac{\Pr(X \text{ is from the claimed speaker})}{\Pr(X \text{ is not from the claimed speaker})} = \frac{\Pr(\lambda_C | X)}{\Pr(\lambda_{\bar{C}} | X)}.$$

If we apply Bayes' rule and discard the constant prior probabilities for claimant and impostor speakers, the likelihood ratio in the log domain becomes

$$\Lambda(X) = \log p(X|\lambda_C) - \log p(X|\lambda_{\bar{C}}).$$

The term $p(X|\lambda_C)$ is the likelihood that the utterance belongs to the claimed speaker and $p(X|\lambda_{\bar{C}})$ is the likelihood that the utterance does not belong to the claimed speaker. The likelihood ratio is compared with a threshold θ and the claimed speaker is accepted if $\Lambda(X) \geq \theta$ and rejected if $\Lambda(X) < \theta$. The likelihood ratio measures how much better the claimant's model scores for the test utterance compared with a nonclaimant model. The decision threshold is then set to adjust the trade-off between rejecting true claimant utterances (false-rejection errors) and accepting nonclaimant utterances (false-acceptance errors). In a real-world application such as telephone banking, this trade-off would be between security and customer satisfaction.

The terms of the likelihood ratio are computed as follows. The likelihood that the utterance X belongs to the claimed speaker is directly computed as

$$\log p(X|\lambda_C) = \frac{1}{T}\sum_{t=1}^{T} \log p(\mathbf{x}|\lambda_C). \qquad (2)$$

The $1/T$ factor is used to normalize the likelihood for utterance duration.

The likelihood that the utterance is not generated from the claimed speaker is formed by using a collection of background-speaker models. With a set of B background-speaker models, $\{\lambda_1, \lambda_2, \ldots, \lambda_B\}$, the background speakers' log-likelihood is computed as

$$\log p(X|\lambda_{\bar{C}}) = \log\left\{\frac{1}{B}\sum_{b=1}^{B} p(X|\lambda_b)\right\},$$

where $p(X|\lambda_b)$ is computed as in Equation 2. Except for the $1/T$ factor, $p(X|\lambda_{\bar{C}})$ is the joint probability density that the utterance comes from a background speaker if we assume equally likely speakers.

Background speakers have been successfully used in several different speaker-verification systems to form various likelihood-ratio tests [10, 21]. The likelihood normalization provided by the background speakers is important for the verification task because it helps minimize the nonspeaker-related variations in the test-utterance scores, allowing stable decision thresholds to be set. The absolute-likelihood score of an utterance from a speaker is influenced by many utterance-dependent factors, including the speaker's vocal characteristics, the linguistic content, and the speech quality. These factors make it difficult to set a decision threshold for absolute-likelihood values to be used during different verification tests. The likelihood-ratio normalization produces a relative score that is more a function of the speaker and less sensitive to nonspeaker utterance variations. Note that the identification task does not need the normalization because decisions are made by using likelihood scores from a single utterance, requiring no inter-utterance likelihood comparisons.

Background-Speaker Selection

Two issues that arise with the use of background speakers for speaker verification are the selection of the speakers and the number of speakers to use. Intuitively, the background speakers should be selected to represent the population of expected impostors, which is generally application specific. In some scenarios, we can assume that impostors will attempt to gain access only from similar-sounding or at least same-sex speakers (dedicated impostors). In a telephone-based application accessible by a larger cross section of potential impostors, on the other hand, the impostors can sound very dissimilar to the users they attack (casual impostors)—for example, a male impostor claiming to be a female user.

Previous systems have relied on selecting background speakers whose models (termed the ratio set, or cohorts) are closest to or most competitive with each enrolled speaker. This choice is appropriate for the dedicated-impostor scenario but, as seen in the experiments and discussed in Reference 10, it leaves the system vulnerable to impostors with very dissimilar voice characteristics. This vulnerability occurs because the dissimilar voice is not well modeled by the numerator or denominator of the likelihood ratio.

Even though we can employ methods of rejecting very dissimilar voices on the basis of thresholding the probability score from the claimed speaker's model [10], the approach of judicious background-speaker selection was pursued here. The experiments that we conducted examine both the same-sex and mixed-sex

impostor situations. Background speakers are selected by using an algorithm described elsewhere [22].

Ideally, the number of background speakers should be as large as possible to model the impostor population better, but practical considerations of computation and storage dictate a small set of background speakers. In the verification experiments, we set the number of background speakers to ten. The limited size was motivated by real-time computation considerations and the desire to set a constant experimental test. For a verification experiment on a given database, each speaker is used as a claimant, while the remaining speakers (excluding the claimant's background speakers) act as impostors and we rotate through all speakers. Large background-speaker sets decrease the number of impostor tests.

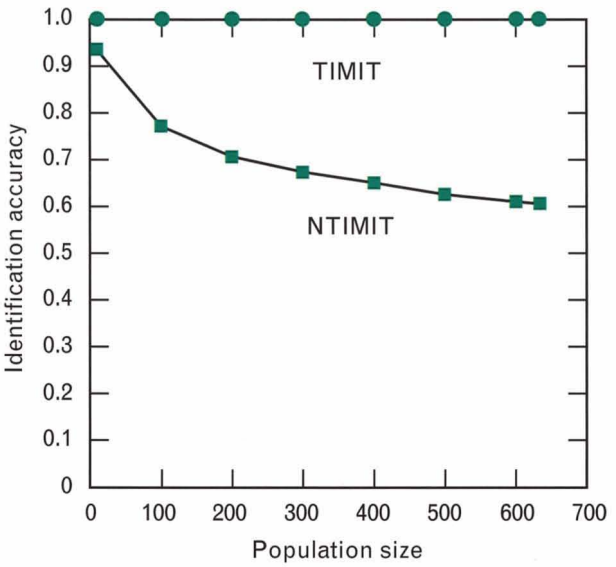


FIGURE 7. Speaker-identification accuracy as a function of population size on the TIMIT and NTIMIT databases. Thirty-two component GMMs were trained with twenty-four seconds of speech and tested with three-second utterances.

Identification Experiments

Identification experiments were conducted on the TIMIT [23], NTIMIT [24], and Switchboard [25] databases (for more information on these databases see the sidebar, "Speaker-Database Descriptions," along with Table 1, which lists the characteristics of each database). The goal of the experiments was to examine the performance of the identification system as a function of population size for both clean wideband speech and telephone speech. The TIMIT performance provides an indication of how crowded the speaker space is under near ideal conditions. The NTIMIT results indicate the performance loss from using noisy telephone speech. Results on the more realistic Switchboard database provide a better measure of expected extemporaneous telephone-speech performance and the effect of handset variability.

TIMIT and NTIMIT Results

For the identification experiments on the TIMIT and NTIMIT databases, all 630 speakers (438 males, 192 females) were used. Speaker models with 32-component GMMs were trained by using eight utterances with a total duration of approximately twenty-four seconds. The remaining two utterances with a duration of approximately three seconds each were individually used as tests (a total of 1260 tests).

Identification accuracy for a population size was computed by performing repeated speaker-identification experiments on fifty sets of speakers randomly selected from the pool of 630 available speakers and averaging the results. This procedure helped average out the bias of a particular population composition. Population sizes of 10, 100, 200, 300, 400, 500, 600, and 630 were used. Figure 7 shows the speaker-identification accuracies for the various populations.

Under the near ideal TIMIT conditions, increasing population size barely affects performance. This result indicates that the limiting factor in speaker-identification performance is not a crowding of the speaker space. However, with telephone-line degradations the NTIMIT accuracy steadily decreases as population size increases. The largest drop in accuracy occurs as the population size increases from 10 to 100. Above 200 speakers the decrease in accuracy becomes almost linear. With the full population of 630 speakers, there is a 39% gap between TIMIT accuracy (99.5%) and NTIMIT accuracy (60.7%). The correct TIMIT speakers have an average rank of 1.01, while the correct NTIMIT speakers have an average rank of 8.29. A speaker's rank for a test utterance is the position of his or her model's score within the sorted list of speaker-model scores, with a rank of 1.0 representing the best-scoring speaker.

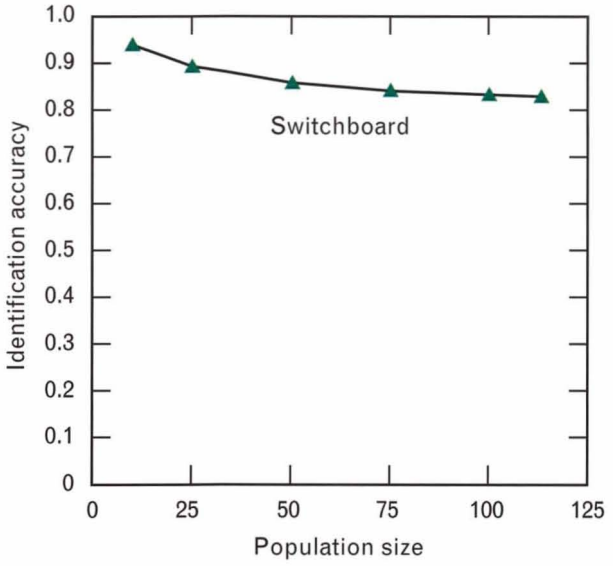


FIGURE 8. Speaker-identification accuracy as a function of population size on the Switchboard database. Sixty-four-component GMMs were trained with six minutes of speech and tested with one-minute utterances.

For the complete 630-population TIMIT database, there are no cross-sex errors, and male and female accuracies are 99.8% and 99.0%, respectively. For the complete 630-population NTIMIT database, there are four cross-sex errors. Accuracy is 62.5% for male speakers versus 56.5% for female speakers.

When we examine the results from the NTIMIT database, the main degradations appear to be noise and bandlimiting. The TIMIT database has an average signal-to-noise ratio (SNR) of 53 dB, while the NTIMIT database has an average SNR of 36 dB. The examination of sweep tones from each telephone line used in the NTIMIT database shows little spectral-shape variability. This result is not surprising, because the telephone handset is the source of most spectral shaping and a single handset was used for all recordings. Detailed studies that systematically impose various degradations on TIMIT speech (e.g., bandlimiting, noise addition) to explain the performance gap between the TIMIT and NTIMIT databases can be found elsewhere [26, 27].

Recently published results based on a different training and testing paradigm with the complete 630-speaker TIMIT database also show a very high accuracy of 95.6% with a text-independent technique that scores only selected phonetic clusters [28]. To the author's knowledge, there have been no published speaker-identification experiments conducted on the complete NTIMIT database.

Switchboard Results

For the Switchboard database, 113 speakers (50 males, 63 females) were used with 64-component GMMs trained by using six minutes of speech extracted equally from two conversations. Testing was performed on a total of 472 utterances of one-minute duration. There were two to twelve test utterances per speaker with an average of four utterances. Identification accuracy was computed as above, except 100 sets per population size were used for populations of 10, 25, 50, 75, 100, and 113. Figure 8 shows the speaker-identification accuracies for the various populations.

Although not directly comparable, the Switchboard results exhibit the same decreasing trend as the NTIMIT results shown in Figure 7, but not as rapidly. Because of the increased training and testing data and the higher SNRs (typically 40 dB or higher), the Switchboard results are higher than the NTIMIT results. For the 113-speaker population, the overall accuracy is 82.8%, with an average rank of 2.29. There are two cross-sex errors, and the male speakers have an accuracy of 81.0% compared with an accuracy of 84.3% for the female speakers.

The effect of handset variability on the results was examined by using the telephone numbers associated with the training and testing utterances. For each conversation in the Switchboard database, a coded version of the callers' telephone numbers was given. Conversations originating from identical telephone numbers were generally assumed to be over the same telephone handset. Conversely, we could have assumed that there is a correlation between conversations originating from different telephone numbers and callers using different handsets.

Neither assumption is strictly true, since callers can use different telephone units with the same telephone number, and similar telephone units can be used at different telephone numbers. There are, of course, other factors, such as different transmission paths and acoustic environments, which also change with different telephone numbers. The aim here was to examine the performance when training and testing utterances

SPEAKER–DATABASE DESCRIPTIONS

FOUR DATABASES WERE USED to conduct speaker-recognition experiments at Lincoln Laboratory: TIMIT, NTIMIT, Switchboard and YOHO (see Table 1).

The TIMIT database, developed by Texas Instruments, Inc. and MIT, allows the examination of speaker-identification performance under almost ideal conditions. With an 8-kHz bandwidth and a lack of intersession variability, acoustic noise, and microphone variability and distortion, TIMIT's recognition errors should be a function of overlapping speaker distributions. Furthermore, each utterance is a read sentence approximately three seconds long. The sentences are designed to contain rich phonetic variability. Because of this variability, speaker-recognition performance that uses three-second TIMIT utterances is higher than using three-second utterances extracted randomly from extemporaneous speech.

The NTIMIT database, developed by NYNEX, is the same speech from the TIMIT database recorded over local and long-distance telephone loops. Each sentence was played through an artificial mouth coupled to a carbon-button telephone handset via a telephone test frame designed to approximate the acoustic coupling between the human mouth and the telephone handset. The speech was transmitted to a local or long-distance central office and looped back for recording.

This arrangement provides the identical TIMIT speech, but degraded through carbon-button transduction and actual telephone line conditions. Performance differences between identical experiments on TIMIT and NTIMIT should arise mainly from the effects of the microphone and telephone transmission degradations.

The Switchboard database, developed by Texas Instruments, Inc., is one of the best telephone-speech, speaker-recognition databases available. Large amounts of spontaneous telephone speech from hundreds of speakers, collected under home and office acoustic conditions with varying telephone handsets, make recognition results from Switchboard more realistic for telephone-based applications. Because the channel conditions tend to be clean, channel noise is not a major issue. However, background noise from radios or televisions can be found in some recordings.

To produce the Switchboard database, engineers recorded each side of a two-way conversation separately to isolate speakers. However, because of performance limits of the telephone-network echo canceling, even single conversation halves may have contained low-level opposite-channel echo. In this work, speaker turns from the transcripts and differential-energy echo suppression were used to isolate single-speaker speech for training and testing.

The YOHO database, developed by ITT, was designed to support text-dependent speaker-verification research such as is used in secure-access technology. It has a well-defined train and test scenario in which each speaker has four enrollment sessions when he or she is prompted to read a series of twenty-four combination-lock phrases. Each phrase is a sequence of three two-digit numbers (e.g., "35-72-41"). There are ten verification trials per speaker, consisting of four phrases per trial. The vocabulary consists of fifty-six two-digit numbers ranging from 21 to 97 (see Reference 10 for the selection rules). The speech was collected in an office environment with a telephone handset connected to a workstation. Thus the speech has a telephone bandwidth of 3.8 kHz, but no telephone transmission degradations.

The YOHO database is different from the above text-independent, telephone-speech databases, which allows us to demonstrate how the GMM verification system, although designed for text-independent operation, can also perform well under the vocabulary-dependent constraints of this application.

Table 1. Characteristics of the Speaker Databases

Database	*Number of Speakers*	*Number of Utterances per Speaker*	*Channel*	*Acoustic Environment*	*Handset*	*Intersession Interval*
TIMIT	630	10 read sentences	Clean	Sound booth	Wideband microphone	None
NTIMIT	630	10 read sentences	PSTN* long distance and local	Sound booth	Fixed carbon button	None
Switchboard	500	1–25 conversation	PSTN long distance	Home and office	Variable	Days-weeks
YOHO	138	4/train, 10/test combination lock	Clean	Office	Telephone, high quality	Days-months

* Public Switched Telephone Network

originate from the same and different telephone numbers under the assumption that the telephone number implies a handset.

Since the speaker models were trained from two conversations, there were at most two training telephone numbers (handsets) per speaker. Of the 113 speakers, 95 trained with utterances from the same telephone number. The first row in Table 2 shows the number of test utterances with and without train/test telephone number matches. A train/test match occurred if a speaker's testing utterance had the same telephone number as either of the training utterances. There is a clear dominance in this test of matched telephone numbers.

The second row of Table 2 shows the number of misclassifications for the two groups. Here we see that most errors are from the mismatched conditions; 45% of the total number of errors come from the mismatched group comprising only 16% of the total number of tests. The error rate of the mismatched group is almost five times that of the matched group, indicating the sensitivity to acoustic mismatches between training and testing conditions. That so many mismatch errors occur even with channel compensation further indicates that the degradations are more complex than a first-order linear filter effect.

Other published speaker-identification results for the Switchboard database typically are based on a smaller 24-speaker set (12 male, 12 female) with a total of 97 test utterances (one to six utterances per speaker). On this task, using ten-second and sixty-second utterances, the GMM system has an accuracy of 94% at ten seconds and 95% at sixty seconds compared with 96% at sixty seconds for ITT's nearest neighbor classifier [29]; 90% at ten seconds and 95% at sixty seconds for BBN's Gaussian classifier [30]; and 89% at ten seconds and 88% at sixty seconds for Dragon Systems' continuous speech-recognition classifier [31]. The testing paradigm was the same for these systems; the training paradigm was not. The accuracy was increased to almost 100% for both of the utterance lengths by using robust scoring techniques [30, 32]. As above, there was significant overlap between training and testing telephone handsets, which favorably biases performance.

Table 2. Switchboard Identification Experiment

	No Matching Telephone Numbers	*Matching Telephone Numbers*
Number of test utterances	74	398
Number of errors	35	43
Percent error	47.3%	10.8%

Verification Experiments

Verification experiments were conducted on the TIMIT, NTIMIT, Switchboard, and YOHO [10, 33] databases. The TIMIT, NTIMIT and Switchboard databases were again used to gauge verification performance over the range of near ideal speech to more realistic, extemporaneous telephone speech. The YOHO database was used to demonstrate performance for a vocabulary-dependent, office-environment, secure-access application. As previously discussed, the composition of the impostor speakers can greatly affect performance. Experiments using same-sex impostors and mixed-sex impostors are presented in conjunction with two different background-speaker selection procedures. There were two same-sex experiments and one mixed-sex experiment: male speakers only (M), female speakers only (F), and male and female speakers together (M+F).

By using the background-speaker selection algorithm [22], we selected two background-speaker sets of size ten from the complete speaker set of each database. In the first background-speaker set, we selected ten speakers who were close to the claimant speaker but maximally spread from each other (denoted 10 msc in the experiments below). In the second background set, we selected five maximally spread close speakers (5 msc) and five speakers who were far from the claimant speaker but maximally spread from each other (5 msf). Since the msf speakers were selected from the complete database, they generally represented opposite-sex speakers. In all experiments, the background speaker's utterances were excluded from the impostor tests.

Results are reported as the equal-error rate (EER) computed by using a global threshold. This threshold is found by placing all the true test scores and impostor test scores in one sorted list and locating the point on the list at which the false acceptance (FA) rate (the percent of impostor tests above the point) equals the false rejection (FR) rate (the percent of true tests below the point); the EER is the FA rate at this point. The EER measures the overall (speaker-independent) system performance by using the largest number of true and impostor tests available.

Results using speaker-dependent thresholds (i.e., treating each speaker's true-utterance and impostor-utterance scores separately) are generally higher than global threshold results, but may have lower statistical significance caused by the use of a smaller number of tests available per speaker.

TIMIT and NTIMIT Results

For the verification experiments on TIMIT and NTIMIT, the 168 speakers (112 males, 56 females) from the test portion of the databases were used. As in the identification experiment, speaker models with 32-component GMMs were trained by using eight utterances with a total duration of approximately twenty-four seconds. The remaining two utterances with duration of approximately three seconds each were individually used as tests. Experiments were performed by using each speaker as a claimant, while the remaining speakers (excluding the claimant's background speakers) acted as impostors, and by rotating through all the other speakers. Table 3 shows the number of claimant and impostor trials for the M, F, and M+F experiments.

Table 3. Claimant and Impostor Trials for the TIMIT and NTIMIT Databases*

Experiment	*Number of speakers*	*Number of true tests per speaker*	*Number of impostor tests per speaker*	*Total number of true tests*	*Total number of impostor tests*
M	112	2	202	224	22,624
F	56	2	88	110	4945
M+F	168	2	313	334	52,538

* Background speaker set size of ten

Table 4. Equal-Error Rate (Percent) for Experiments on the TIMIT and NTIMIT Databases*

Database	*M (10 msc)*	*M (5 msc, 5 msf)*	*F (10 msc)*	*F (5 msc, 5 msf)*	*M+F (10 msc)*	*M+F (5 msc, 5 msf)*
TIMIT	0.14	0.32	0.28	0.71	0.50	0.24
NTIMIT	8.15	8.48	8.79	10.44	8.68	7.19

* msc indicates maximally spread close-background speakers; msf indicates maximally spread far-background speakers

Table 4 shows the results for the three experimental conditions (M, F, and M+F) and two background-speaker selections. As with the speaker-identification results, almost perfect performance is obtained on the TIMIT database; the NTIMIT performance is significantly worse. The NTIMIT best M+F EER is about thirty times worse than the TIMIT M+F EER. Comparing the M+F experiments with and without the far-background speakers makes it clear that inclusion of the dissimilar speakers improved performance by better modeling the impostor population. As expected, the dissimilar speakers for the male speakers were mainly female speakers, and vice versa. However, since there was a predominance of male speakers in the M+F test, the improvement is not as great as may have occurred with a more balanced test.

Switchboard Results

The verification paradigm on the Switchboard database was different from that used on the TIMIT and NTIMIT databases. Here, 24 claimant speakers (12 males, 12 females) were each represented by 64-component GMMs trained by using three minutes of speech extracted equally from four conversations. A total of 97 claimant utterances of sixteen-second average duration were selected from conversations. Claimants had between one and six true tests with an average of four. A separate set of 428 utterances of sixteen-second average duration from 210 speakers (99 males and 111 females) was used for the impostor tests. The utterances were designated by using speaker turns from the transcripts to isolate single-speaker speech. Table 5 shows the number of claimant and impostor trials for the M, F, and M+F experiments.

Two background-speaker sets were used from this relatively small claimant population: a same-sex set (ss), in which each speaker used all other claimant speakers of the same sex as background speakers, and a selection consisting of five maximally spread close-background and five maximally spread far-background speakers (essentially a mixed-sex set). Table 6 shows the results for these experiments.

We were initially surprised to see that the same-sex background set (11 ss) did worse than the mixed-sex background set (5 msc, 5 msf) on the M and F experiments. Since same-sex impostors were used in

Table 5. Claimant and Impostor Trials for the Switchboard Database*

Experiment	*Number of speakers*	*Average number of true tests per speaker*	*Number of impostor tests per speaker*	*Total number of true tests*	*Total number of impostor tests*
M	12	4	210	47	2520
F	12	4	218	50	2616
M+F	24	4	428	97	10,272

* Separate claimant and impostor populations used

Table 6. Equal-Error Rate (Percent) for Experiments on the Switchboard Database*

Database	*M (11 ss)*	*M (5 msc, 5 msf)*	*F (11 ss)*	*F (5 msc, 5 msf)*	*M+F (11 ss)*	*M+F (5 msc, 5 msf)*
Switchboard	5.83	4.25	11.39	7.99	8.25	5.15

* msc indicates maximally spread close-background speakers; msf indicates maximally spread far-background speakers; ss indicates same sex

these tests, we expected that using same-sex background speakers would perform better than a background set split between males and females.

However, closer examination of the utterances in error showed that they generally were extracted from a mixed-sex conversation and that the echo from the opposite side was contaminating the utterance. Thus, for example, some ostensibly male-only impostor utterances actually contained female speech. As with the TIMIT and NTIMIT experiments, a decrease in EER was obtained in the M+F experiment by using the mixed sex (close and far) background-speaker set.

Examination of the claimant-training and claimant-testing utterance telephone numbers also found that only sixteen of the claimant tests were from telephone numbers unseen in the training data, which favorably biases the FR rate. In the mismatched cases, some speakers had high FR errors.

YOHO Results

For the YOHO experiments, each speaker was modeled by a 64-component GMM trained by using the four enrollment sessions (average of six minutes). Each speaker had ten verification sessions consisting of four combination-lock phrases (average of fifteen seconds). Experiments consisted of using each speaker as a claimant, while the remaining speakers (excluding the claimant's background speakers) acted as impostors, and rotating through all speakers. Like the TIMIT and NTIMIT databases, there was a gender imbalance: 106 male speaker and 32 female speakers. Table 7 displays the number of claimant and impostor trials for the M, F, and M+F experiments.

Table 8 gives results for three experimental conditions with the two background-speaker sets. In addition to the EER, the table also gives the false-rejection rate at false-acceptance rates of 0.1% and 0.01%. These latter numbers measure performance at tight operating specification for an access-control application. We see that very low error rates are achievable for this task because of the good quality and vocabulary constraints of the speech. The vocabulary constraints mean that a speaker's GMM need model only a constrained acoustic space, thus allowing an inherently text-independent model to use the text-dependent training and testing data effectively.

The high performance is also found for identification using the same data: accuracies of 99.7% for

Table 7. Claimant and Impostor Trials for the YOHO Database*

Experiment	*Number of speakers*	*Number of true tests per speaker*	*Number of impostor tests per speaker*	*Total number of true tests*	*Total number of impostor tests*
M	106	10	950	1060	100,700
F	32	10	210	318	6720
M+F	138	10	1268	1378	175,105

* Background speaker set size of ten

Table 8. Equal-Error Rate (Percent) and False-Rejection Rate at False-Acceptance Rates of 0.1% and 0.01% for Experiments on the YOHO Database*

Database	*M (10 msc)*	*M (5 msc, 5 msf)*	*F (10 msc)*	*F (5 msc, 5 msf)*	*M+F (10 msc)*	*M+F (5 msc, 5 msf)*
YOHO	0.20	0.28	1.88	1.57	0.58	0.51
FR at FA = 0.1%	0.38	0.38	1.89	1.89	0.87	0.65
FR at FA = 0.01%	0.94	2.36	2.51	3.77	2.40	2.40

* msc indicates maximally spread close-background speakers; msf indicates maximally spread far-background speakers

males, 97.8% for females, and 99.3% for males and females. The close-background and far-background selections boosted performance for the M+F experiment, which again was dominated by male speakers.

J.P. Campbell presents verification and identification results on the YOHO database from several different systems [33]. Compared with the 0.5% EER of the GMM system, ITT's continuous speech-recognition classifier has an EER of 1.7% [10], ITT's nearest neighbor classifier has an EER of 0.5%, and Rutgers University's neural tree network has an EER of 0.7% [34]. These results can be only loosely compared, however, since different training and testing paradigms and background speaker sets were used (e.g., ITT's continuous speech-recognition system uses five background speakers).

Conclusion

In this article, we have reviewed the research, development, and evaluation of automatic speaker-recognition systems at Lincoln Laboratory. Starting from the speaker-dependent vocal-tract information conveyed via the speech spectrum, we outlined the development of a statistical speaker-model approach to represent the underlying characteristic vocal-tract shapes of a person's voice. With a text-independent assumption, this statistical speaker model leads to the Gaussian mixture speaker model that serves as the basis for our speaker identification and verification systems. The Gaussian mixture model provides a simple yet effective speaker representation that is computationally inexpensive and provides high recognition accuracy on a wide range of speaker recognition tasks.

Experimental evaluation of the performance of the automatic speaker-recognition systems was conducted on four publicly available speech databases: TIMIT, NTIMIT, Switchboard, and YOHO. Each database offers different levels of speech quality and control. The TIMIT database provides near ideal speech with high-quality clean wideband recordings, no intersession variabilities, and phonetically rich read speech. Under these ideal conditions, we determined that crowding of the speaker space was not an issue for population sizes up to 630. An identification accuracy of 99.5% was achieved for the complete 630-speaker population. The NTIMIT database adds real telephone line degradations to the TIMIT data, and these degradations caused large performance losses. The NTIMIT accuracy dropped to 60.7% for the same 630-population identification task. For verification, the TIMIT EER was 0.24%, compared with 7.19% on NTIMIT.

The Switchboard database provides the most realistic mix of real-world variabilities that can affect speaker-recognition performance. The performance trends on Switchboard appeared similar to those found with NTIMIT, producing an 82.8% identification accuracy for a 113-speaker population and an EER of 5.15% for a 24-speaker verification experiment. The factors degrading the NTIMIT and Switchboard performances, however, are different. High noise levels seem to be the main degradation in NTIMIT, whereas handset variability and cross-channel echo are the two major degradations in Switchboard. For the identification experiments, we found that the error rate for utterances from telephone

numbers unseen in the training utterances was almost five times that of utterances from telephone numbers found in the training utterances.

Finally, results on the YOHO database show that low error rates are possible for a secure-access verification application even with a text-independent verification system. An overall EER of 0.51% and a false-rejection rate of 0.65% at a 0.1% false-acceptance rate were obtained. The constrained vocabulary along with the good-quality speech allowed the model to focus on the sounds that characterize a person's voice without extraneous channel variabilities.

As the experimental results show, speaker-recognition performance is indeed at a usable level for particular tasks such as access-control authentication. The major limiting factor under less controlled situations is the lack of robustness to transmission degradations, such as noise and microphone variabilities. Large efforts are under way to address these limitations, exploring areas such as understanding and modeling the effects of degradations on spectral features, applying more sophisticated channel compensation techniques, and searching for features more immune to channel degradations.

For Further Reading

Most current research in speaker-recognition systems is published in the proceedings from the following conferences: International Conference on Acoustics, Speech and Signal Processing (ICASSP), International Conference on Spoken Language Processing (ICSLP), and European Conference on Speech Communication and Technology (Eurospeech). Other publications that feature speaker-recognition research are *IEEE Transactions on Speech and Audio Processing* and *ESCA Speech Communication Journal.* Excellent, general review articles on the area of speaker recognition can be found in References 3 and 35 through 38.

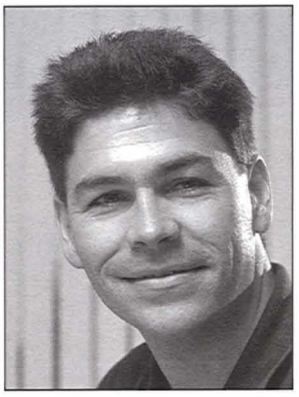
Acknowledgments

The author wishes to thank Beth Carlson, Richard Lippmann, Jerry O'Leary, Doug Paul, Cliff Weinstein, and Marc Zissman of the Speech Systems Technology group for many helpful technical discussions and assistance throughout this work.

REFERENCES

1. F. McGehee, "The Reliability of the Identification of the Human Voice," *J. General Psychology* **17**, 249 (1937)
2. W. D. Voiers, "Perceptual Bases of Speaker Identity," *J. Acoust. Soc. Am.* **36**, 1065 (1964).
3. G.R. Doddington, "Speaker Recognition—Identifying People by Their Voices," *Proc. IEEE* **73**, 1651 (1985).
4. A.E. Rosenberg, "Listener Performance in Speaker-Verification Tasks," *IEEE Trans. Audio Electroacoust.* **AU-21**, 221 (1973).
5. D.A. Reynolds and L.P. Heck, "Integration of Speaker and Speech Recognition Systems," *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing* **2**, *Toronto, 14–17 May 1991,* p. 869.
6. C. Schmandt and B. Arons, "A Conversational Telephone Messaging System," *IEEE Trans. Consum. Electron.* **CE-30**, xxi (Aug. 1984).
7. L. Wilcox, F. Chen, D. Kimber, and V. Balasubramanian, "Segmentation of Speech Using Speaker Identification," *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, Adelaide, Australia, 19–22 Apr. 1994,* p. I-161.
8. B.M. Arons, "Interactively Skimming Recorded Speech," Ph.D. Thesis, MIT, Cambridge, MA, 1994.
9. J.M. Naik and G.R. Doddington, "Evaluation of a High Performance Speaker-Verification System for Access Control," *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing* **4**, *Dallas, 6–9 Apr. 1987,* p. 2392.
10. A. Higgins, L. Bahler, and J. Porter, "Speaker Verification Using Randomized Phrase Prompting," *Digital Signal Process.* **1**, 89 (1991).
11. D.A. Reynolds, "A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification," Ph.D. Thesis, Georgia Institute of Technology, Atlanta, GA, 1992.
12. D.A. Reynolds, R.C. Rose, and M.J.T. Smith, "PC-Based TMS320C30 Implementation of the Gaussian Mixture Model Text-Independent Speaker-Recognition System," *Proc. Int. Conf. on Signal Processing Applications and Technology* **2**, *Boston, 2–5 Nov. 1992,* p. 967.
13. S.B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-28**, 357 (1980).
14. D.A. Reynolds, "Experimental Evaluation of Features for Robust Speaker Identification," *IEEE Trans. Speech Audio Process.* **2**, 639 (1994).
15. D.B. Paul, "Speech Recognition Using Hidden Markov Models," *Linc. Lab. J.* **3**, 41 (1990).
16. M.A. Zissman, "Automatic Language Identification of Telephone Speech," in this issue.
17. L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE* **77**, 257 (1989).
18. N.Z. Tishby, "On the Application of Mixture AR Hidden Markov Models to Text Independent Speaker Recognition," *IEEE Trans. Signal Process.* **39**, 563 (1991).
19. D.A. Reynolds and R.C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Trans. Speech Audio Process.* **3**, 72 (1995).
20. A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc.* **39**, 1 (1977).

21. A.E. Rosenberg, J. DeLong, C.H. Lee, B.H. Juang, and F.K. Soong, "The Use of Cohort Normalized Scores for Speaker Verification," *Int. Conf. on Speech and Language Processing* **1**, *Banff, Alberta, Canada, 12–16 Oct. 1992,* p. 599.
22. D.A. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models," *Speech Commun.* **17**, 91 (Aug. 1995).
23. W.M. Fisher, G.R. Doddington, and K.M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," *Proc. DARPA Workshop on Speech Recognition, Palo Alto, CA, Feb. 1986,* p. 93.
24. C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A Phonetically Balanced, Continuous Speech Telephone Bandwidth Speech Database," *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing* **1**, *Albuquerque, 3–6 Apr. 1990,* p. 109.
25. J.J. Godfrey, E.C. Holliman, and J. MacDaniel, "Switchboard: Telephone Speech Corpus for Research and Development," *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing* **1**, *San Francisco, 23–26 Mar. 1992,* p. I-517.
26. D.A. Reynolds, "Large Population Speaker Recognition Using Wideband and Telephone Speech," *SPIE* **2277**, 11 (1994).
27. D.A. Reynolds, M.A. Zissman, T.F. Quatieri, G.C. O'Leary, and B.A. Carlson, "The Effects of Telephone Transmission Degradations on Speaker Recognition Performance," *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing* **1**, *Detroit, 9–12 May 1995,* p. 329.
28. J.-L. Floch, C. Montacié, and M.-J. Caraty, "Investigations on Speaker Characterization from Orphée System Technics," *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, Adelaide, Australia, 19–22 Apr. 1994,* p. I-149.
29. A.L. Higgins, L.G. Bahler, and J.E. Porter, "Voice Identification Using Nearest-Neighbor Distance Measure," *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, Minneapolis, 27–30 Apr. 1993,* p. II-375.
30. H. Gish and M. Schmidt, "Text-Independent Speaker Identification," *IEEE Signal Process. Mag.* **11**, 8 (Oct. 1994).
31. L. Gillick, J. Baker, J. Baker, J. Bridle, M. Hunt, Y. Ito, S. Lowe, J. Orloff, B. Peskin, R. Roth, and F. Scallone, "Application of Large Vocabulary Continuous Speech Recognition to Topic and Speaker Identification Using Telephone Speech," *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, Minneapolis, 27–30 Apr. 1993,* p. II-471.
32. L.G. Bahler, J.E. Porter, and A.L. Higgins, "Improved Voice Identification Using a Nearest-Neighbor Distance Measure," *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, Adelaide, Australia, 19–22 Apr. 1994,* p. I-321.
33. J.P. Campbell, Jr., "Testing with the YOHO CD-ROM Voice Verification Corpus," *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, Detroit* **1**, *9–12 May 1995,* p. 341.
34. H.-S. Liou and R. Mammone, "A Subword Neural Tree Network Approach to Text-Dependent Speaker Verification," *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing* **1**, *Detroit, 9–12 May 1995,* p. 357.
35. B.S. Atal, "Automatic Recognition of Speakers from Their Voices," *Proc. IEEE* **64**, 460 (1976).
36. A.E. Rosenberg, "Automatic Speaker Verification: A Review," *Proc. IEEE* **64**, 475 (1976).
37. D. O'Shaughnessy, "Speaker Recognition," *IEEE ASSP Mag.* **3**, 4 (Oct. 1986).
38. J.M. Naik, "Speaker Verification: A Tutorial," *IEEE Commun. Mag.* **28**, 42 (Jan. 1990).



DOUGLAS A. REYNOLDS
is a staff member in the Speech Systems Technology group. He received his B.E.E. and Ph.D. degrees from the School of Electrical Engineering at the Georgia Institute of Technology. Doug worked as a summer staff member in the Speech Systems Technology group in 1989 and 1991 before joining the group full time as a staff member in 1992. His research focus is on robust speaker recognition, robust processing for degraded speech recognition, and applications of speaker verification for secure-access control.