Speaker Recognition—Identifying People by their Voices

GEORGE R. DODDINGTON, MEMBER, IEEE

Invited Paper

The usefulness of identifying a person from the characteristics of his voice is increasing with the growing importance of automatic information processing and telecommunications. This paper reviews the voice characteristics and identification techniques used in recognizing people by their voices. A discussion of inherent performance limitations, along with a review of the performance achieved by listening, visual examination of spectrograms, and automatic computer techniques, attempts to provide a perspective with which to evaluate the potential of speaker recognition and productive directions for research into and application of speaker recognition technology.

I. INTRODUCTION

The human ear is a marvelous organ. Beyond our uniquely human ability to receive and decode spoken language, the ear supplies us with the ability to perform many diverse functions. These include, for example, localization of objects, enjoyment of music, and the identification of people by their voices. Currently, along with efforts to develop computer procedures that understand spoken messages, there is also considerable interest in developing procedures that identify people from their voices. The purpose of this paper is to review this speaker recognition problem and the technology being developed and applied to solve it. First, however, it might be appropriate to discuss the motivation for such study. Why develop a speaker recognition machine?

Speaker recognition is an example of biometric personal identification. This term is used to differentiate techniques that base identification on certain intrinsic characteristics of the person (such as voice, fingerprints, retinal patterns, or genetic structure) from those that use artifacts for identification (such as keys, badges, magnetic cards, or memorized passwords). This distinction confers upon biometric techniques the implication of greater identification reliability, perhaps even infallibility, because the intrinsic biometrics are presumed to be more reliable than artifacts, perhaps even unique. Thus a prime motivation for studying speaker recognition is to achieve more reliable personal identification. This is particularly true for security applications, such as physical access control (a voice-actuated door lock for your home or ignition switch for your automobile), computer data access control, or automatic telephone transaction control (airline reservations or bank-by-phone). Convenience is another benefit which accrues to a biometric system, since biometric attributes cannot be lost or forgotten and thus need not be remembered.

Applications also exist which depend uniquely upon the identification of a person by his voice. Such applications include forensic science and the automated processing of reconnaissance information. For example, 32 channels of enemy air-to-ground telecommunications are being monitored to detect activities of the Red Baron. Is he in the air now? Or an axe murderer telephones the location of his victim's body to the police. Does the suspect's voice match the murderer's? The identification problems posed by these applications can only be solved by speaker recognition technology. So far, all of the applications that have been mentioned fall into a category which may be called voice verification. But there are several different speaker recognition task definitions, with different performance characteristics for each. These will now be described.

A. Types of Speaker Recognition Tasks and Applications

Speaker recognition is a generic term which refers to any task which discriminates between people based upon their voice characteristics. Within this general task description there are two specific tasks that have been studied extensively. These are referred to as speaker identification and speaker verification. (Sometimes the term "voice" or "talker" is substituted for "speaker," and sometimes the term "authentication" is substituted for "verification." Thus for example, speaker verification and voice authentication refer to the same task.) The distinction between identification and verification is simple: The speaker identification task is to classify an unlabeled voice token as belonging to (having been spoken by) one of a set of N reference speakers (N possible outcomes), whereas the speaker verification task is to decide whether or not an unlabeled voice token belongs to a specific reference speaker (2 possible outcomes—the token is either accepted as belonging to the reference speaker or is rejected as belonging to an

Manuscript received March 20, 1985; revised July 18, 1985.

The author is with Speech Research, Computer Science Laboratories, Texas Instruments Inc., Dallas, TX 75265, USA.

0018-9219/85/1100-1651$01.00

impostor). Note that the information in bits, denoted I, to be gained from the identification task is in general greater than that to be gained from the verification task:

$I_{ident} = \log_2(N)$ (assuming equal *a priori* probability of occurrence for all reference speakers)

$I_{ver} = 1$ (assuming *a priori* probability of occurence of reference speaker = 0.5).

It is natural then to expect that, all other factors being equal, recognition performance (i.e., probability of error) will be better for the verification task than for the identification task. An example of this contrast is shown in Fig. 1

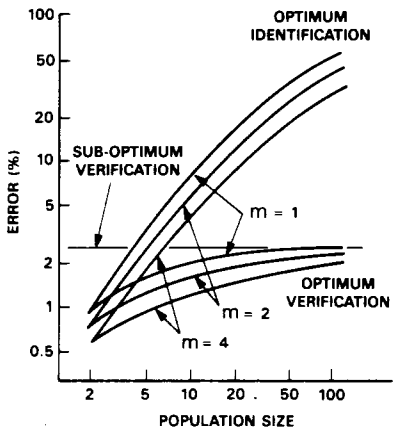

Fig. 1. This figure illustrates the relative performance for speaker verification and speaker identification for a hypothetical task in which the feature vector for speaker i is distributed as a multivariate Gaussian distribution with mean vector $u(i)$ and an identity covariance matrix and further in which the speakers are chosen randomly so that the mean vector for speaker i is also a multivariate Gaussian distribution with zero mean and a covariance matrix equal to a constant S times the identity matrix [19]. For the identification task, the decision is to choose that speaker for which the likelihood of the input vector is maximum, given the observed feature vector. For the verification task there are three possible decisions. First, the optimum decision is to choose that decision hypothesis for which the likelihood of the observed feature vector is greatest, given knowledge of the impostor distributions. Second, since it is unlikely that there will be knowledge of the impostors in any real application, the distribution of the impostor population is estimated for indefinitely large impostor populations. Finally, the easiest and often a completely satisfactory approach is to assume that the probability distribution for impostors is extremely diffuse and therefore to assume a constant likelihood function for impostors. The results of a computer simulation of this example are plotted for three values of speech feature vector dimensionality; namely, $m = 1$, 2, and 4. The variance of impostor mean features S was chosen to provide a suboptimum verification error rate of 2.5 percent. ($S = 2557$ for $m = 1$, $S = 72.4$ for $m = 2$, and $S = 11.0$ for $m = 4$.) The various recognition hypotheses were assigned equal *a priori* likelihood, and the error rates shown were obtained by averaging over a very large number of randomly chosen populations. Note that optimum verification always yields better error rates than optimum identification, except when the population size is 2, in which case verification and identification are the same task. The most important result illustrated by this example is that verification performance, even for suboptimum decision rules, remains satisfactory as the population size becomes large, whereas the performance for speaker identification continues to degrade with the identification error rate approaching 100 percent for sufficiently large populations.

for optimum decision rules applied to a multivariate Gaussian model. In fact, population size is a critical performance parameter for speaker identification, with the probability of error approaching 1 for indefinitely large populations. The performance for speaker verification is unaffected by population size, however. Although the performance for speaker verification is stable with increasing speaker population size, there is one difficulty with verification that is not present with identification. This is, namely, that a much more comprehensive grasp of the variability of the speech features used for discrimination is required in the verification task. Thus while determining which reference is "closest" to the input token may serve the identification task reasonably well without statistical calibration, the verification task demands proper statistical characterization of verification features in order to judge "close enough."

Speaker identification as defined above is also sometimes called "closed-set" identification, which contrasts it from "open-set" identification. In open-set identification the possibility exists that the unknown voice token does not belong to any of the reference speakers. The number of possible decisions is then $N + 1$, which includes the option to declare that the unknown token belongs to none of the reference speakers. Thus open-set identification is a combination of the identification and verification tasks which combines the worst of both—performance is degraded by the complexity of the identification task, and the rejection option requires good characterization of speech feature statistics.

The degree of control over the generation of the speech token is another important speaker recognition parameter. Fixed-text or text-dependent speech tokens are used in applications in which the unknown speaker wishes to be recognized and is therefore cooperative. Free-text or text-independent speech tokens are required in those applications where such control cannot be maintained, either because the speaker is not cooperative or perhaps because the recognition must be done unobtrusively. Generally speaking, recognition performance in fixed-text applications will be better than in free-text applications, because better calibration of the input speech token is possible with identical reference speech material and because the ability to control the text of the input speech will often extend also to the control of the speaker and his environment. Serving to offset this somewhat, free-text applications often provide longer samples of speech for recognition, which tends to improve recognition performance.

Armed with these task definitions we can now tabulate the recognition tasks required by the various applications. This is done in Table 1. Of these applications, security applications will exhibit the best speaker recognition performance because of the cooperative user and controlled conditions. A most difficult problem in the forensic and reconnaissance applications is to establish a valid statistical model upon which verification decisions may be based. An effective statistical model is difficult to establish in this environment because of the lack of control over the speech signal and the speaker and also because of the difficulty in predicting acoustical and transmission conditions. Even with an adequate statistical model for making verification decisions, the lack of control in text-independent applications invariably results in much poorer recognition performance than for those applications in which control is exercised over all aspects of the speaker verification task.

Table 1 A Tabulation of the Speaker Recognition Tasks Required by Various Applications

		IDENTIFICATION		VERIFICATION	
		open-set	closed-set	free-text	fixed-text
SECURITY:	Physical entry				×
	Database access				×
	Telephone transactions			×	×
RECONNAISSANCE		×			
FORENSIC APPLICATIONS				×	×

Note that almost all applications are classified as verification tasks in Table 1. Even the reconnaissance task can be viewed as a set of verification decisions. It is difficult for me to visualize a real operational application of speaker identification. Yet the identification task formulation remains popular in laboratory evaluations.

B. Inherent Factors which Limit the Recognizability of Speakers

What is it about the speech signal that conveys information about the speaker's identity? There are, of course, many different sources of speaker identifying information, including high-level information such as dialect, subject matter or context, and style of speech (including lexical and syntactical patterns of usage). This high-level information is certainly valuable as an aid to recognition of speakers by human listeners, but it has not been used in automatic recognition systems because of practical difficulties in acquiring and using such information. Rather, automatic techniques focus on "low-level" acoustical features. These low-level features include such characteristics of the speech signal as spectral amplitudes, voice pitch frequency, formant frequencies and bandwidths, and characteristic voicing aperiodicities. (See [22] for a good review of such features.) These variables may be measured as a function of time or the statistics of long-term averages may be used as recognition variables. But the real question, the essence of the problem, is this: How stable are these speaker discriminating features? Given a speech signal, is the identity of the speaker uniquely decodable?

The fact is that the speech signal is a complex function of the speaker and his environment. It is an acoustic signal generated by the speaker and which does not convey detailed anatomical information, at least not in any explicit manner. This distinguishes voice recognition from fingerprint identification, since fingerprint recognition (along with other physical attributes such as hand geometry and retinal patterns) uses fixed, static, physical characteristics, while speaker recognition (along with signature recognition) uses dynamic "performance" features that depend upon an act.

Thus there exist inherent limitations in performance which are attributable to the nature of the speech signal and its relationship to the signal generator (the speaker). To appreciate these limits we must understand the source of speaker-discriminating information and how it is encoded in the speech signal. The speech signal, being a consequence of articulation, is determined by the vocal apparatus and its neural control. Thus there are two possible sources of speaker information; namely, the physical and structural characteristics of the vocal tract and the controlling information from the brain and articulatory masculature. This information is imparted to the speech signal during articulation along with all the other information sources. These other sources include not only the linguistic message but also the speech effort level (loud, soft), emotional state (e.g., anger, fear, urgency), health, age, and so on.

The characteristics of the speech signal are determined primarily by the linguistic message, via control of the vocal tract masculature and the resulting articulation of the vocal cords, jaw, tongue, lips, and velum (which controls coupling to the nasal cavity). This articulation, in turn, produces the speech signal as a complex function of the articulatory parameters. The secondary speech messages, including speaker discriminants, are encoded as nonlinguistic articulatory variations of the basic linguistic message. Thus the information useful for identifying the speaker is carried indirectly in the speech signal, a side effect of the articulatory process, and the speaker information may be viewed as "noise" applied to the basic linguistic message. Thus the problem with speaker recognition is that there are no known speech features or feature transformations which are dedicated solely to carrying speaker-discriminating information, and further that the speaker-discriminating information is a second-order effect in the speech features.

The fact is, however, that different individuals typically exhibit speech signal characteristics that are quite strikingly individualistic. We know that people sound different from each other, but the differences become visually apparent when comparing spectrograms from different individuals. The spectrogram is by far the most popular and generally informative tool available for phonetic analysis of speech signals. The spectrogram is a running display of the spectral amplitude of a short-time spectrum as a function of frequency and time. The amplitude is only rather crudely plotted as the level of darkness, but the resonant frequencies of the vocal tract are usually clearly represented in the spectrogram. Fig. 2 demonstrates the degree of difference between spectrograms of five different men saying "Berlin Forest." Note the differences between the individual renditions of this linguistic message. Segment durations, formant frequencies, and formant frequency transitions, pitch and pitch dynamics, formant amplitudes, all exhibit gross differences from speaker to speaker. Thus these speakers would be very easy to discriminate by visual inspection of their spectrograms. This is especially impressive in view of the fact that two of the speakers illustrated in Fig. 2 are identical twins who sound quite similar to each other. Yet their spectrograms look very different.

But there are problems with this appealing notion of spectrographic differences. The primary difficulty lies not with the similarity between different speakers. Speakers usually sound very different from each other, and, in fact, the spectrograms in Fig. 2 show large differences between speakers. The real problem is that a single speaker also

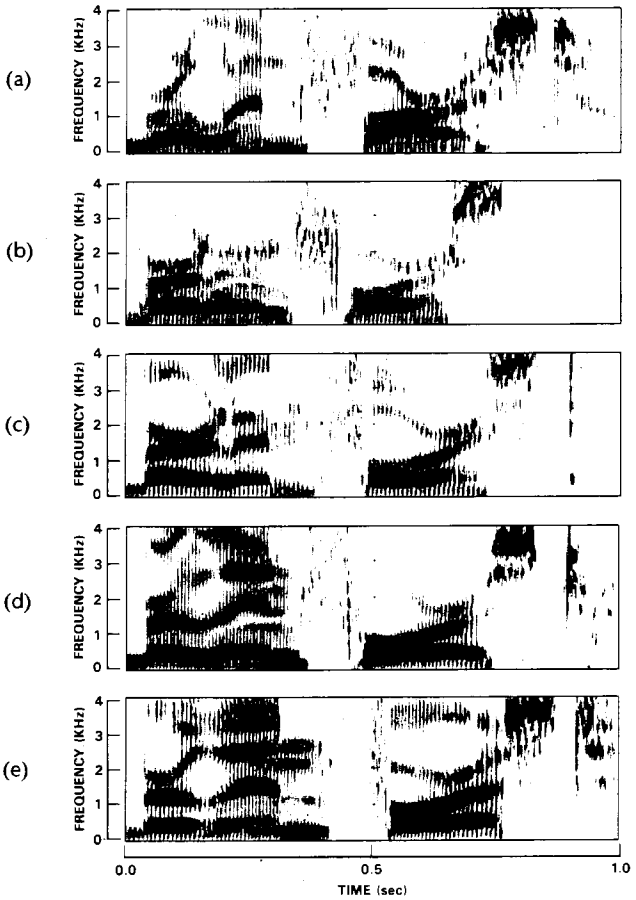

Fig. 2. This figure exhibits five different spectrograms, one each from five different men. The spectrogram is a display of the amplitude of a speech signal as a function of frequency and time. The spectral amplitude is computed as a short-time Fourier transform of the speech signal and is plotted on the *z*-axis, with greater spectral amplitude being depicted as a darker marking. The running window used in the spectral analysis is 6 ms long and the signal is weighted by a Hamming window. The abscissa is the time axis, with 1 s being displayed in this figure. The ordinate is the frequency axis, which spans the range from 0 Hz to 4 kHz in this figure. Although the spectrogram represents amplitude only imprecisely, a great deal of phonetic information may be decoded from the spectrogram by an expert phonetician. Indeed, many studies have shown that the energy loci in frequency and time are perceptually more important than exact calibration of the spectral amplitudes. (These energy loci are usually referred to as "formant" frequencies by phoneticians.) The utterance spoken in each of the five speech spectrograms displayed in this figure is "Berlin Forest." Note that although there are general similarities in the spectrograms as dictated by the linguistic message, the speaker differences are striking. So striking, in fact, that one might be tempted to question the sameness of the phonetic transcription. Two of the speaker represented in this figure are identical twins, and their voices do sound quite similar. These are speakers (b) and (c). Even for these twins, the spectrograms are unquestionably different.

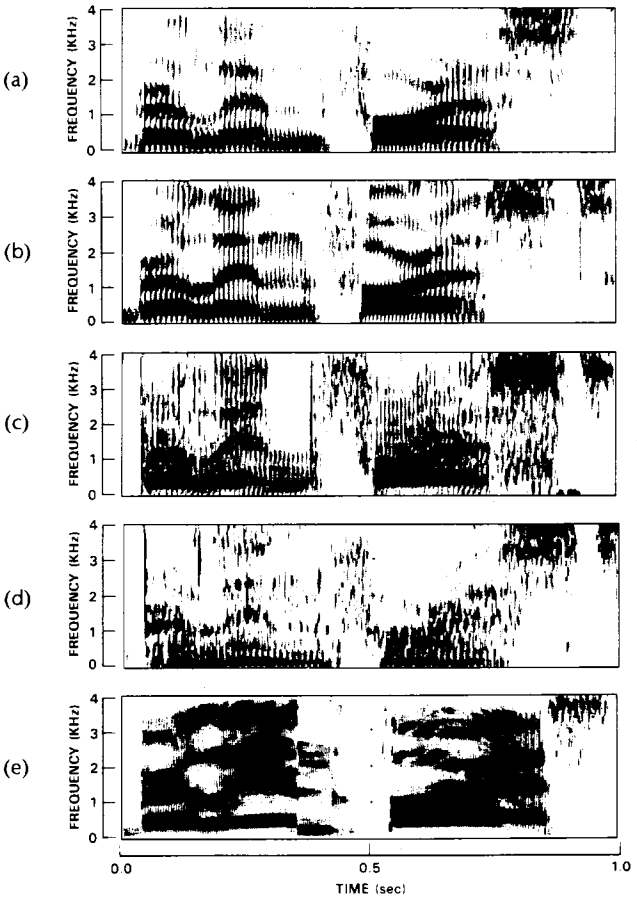

Fig. 3. This figure exhibits five different speech spectrograms, all from the same speaker, one who is not represented in Fig. 2. The utterance spoken in these spectrograms is the same as in Fig. 2. Note that the variation in spectrograms produced by the same person can be extremely great, even greater than the differences between speakers seen in Fig. 2. Examples (a) and (b) are for speech under nominal conditions, but taken from two different recording sessions. There are some significant amplitude differences, particularly above 2 kHz, but the formant frequencies remain nearly identical in frequency and time. Example (c) is for speech collected through a carbon button microphone. Notice that the nonlinearities of this microphone create significant spectral distortions and that the weaker formant frequencies above 21 kHz are largely obscured. Example (d) is for very softly spoken speech (about 20 dB below a normal comfortable level). Notice several changes: First, the spectrum falls off more rapidly with frequency, with most of the signal energy appearing below 1 kHz. Second, the spectrum appears "noisy," which is largely attributable to irregular voiced excitation. These changes also make the formant frequencies much less distinct. Example (e) is for very loudly spoken speech (about 20 dB above a normal comfortable level). The spectrogram of this speech signal bears little resemblance to that in (a) (at least relative to the expected between-speaker differences exhibited in Fig. 2), with a much higher pitch frequency and with relatively more energy at the higher frequencies. Indeed, the two speech signals do not sound much like the same person, either.

often sounds (and looks, spectrographically) very different from time to time. The problem is not so much telling people apart. Rather, the problem is that people sometimes are just not themselves(!) We call this phenomenon "intra-speaker variability." This is illustrated in Fig. 3, which displays the spectrograms of five different tokens (of the same words, "Berlin Forest") for one single speaker. (This speaker is distinct from those of Fig. 2.) Note that the differences between spectrogram (a) and spectrogram (b) are quite small, but that for other conditions the differences between spectrograms are quite marked. Even under the same experimental conditions, a significant difference between spectrograms may be observed for speech data collected at

different times. Differences attributable to significant changes in either the transmission path (carbon button microphone substituted for the linear dynamic microphone) or speaker variation (sotto voce or shouted speech) may become so large as to render any speaker recognition decision completely unreliable.

II. The Acoustical Bases for Speaker Recognition

Before addressing the various speaker recognition technologies and systems, it is appropriate to review the acoustical bases for speaker recognition. An excellent exposition on this subject is presented in a book by Francis Nolan [36]; see also [12]. Note that from introspection of our own personal experience we can appreciate a wide variety of nonacoustic clues that may be used unconsciously to aid the identification of a speaker. Although we usually think of speaker information as comprising acoustical attributes and related perceptual features such as "breathiness" or "nasality," there are other perhaps more useful characteristics that we as listeners use all the time. Such sources include speaker dialect, style of speech (for example, energetic, sexy, sarcastic, witty), and particular unique verbal mannerisms (for example, use of particular words and idioms, or a particular kind of laugh). Our interest here, however, is principally in the low-level acoustical characteristics of the speech signal which support speaker recognition, rather than the higher level features mentioned above. Lack of interest in these higher level sources of information stems from a variety of issues, including the need for extensive training (personal knowledge of the speaker), difficulty in quantifying these features, and the resultant difficulty in controlling and evaluating performance as a function of this information.

One seemingly reasonable way to approach the task of determining speaker-characterizing acoustical features is to determine the perceptual bases for speaker recognition and then to determine the acoustical correlates of these perceptual features [5], [10]. Unfortunately, this method has not been productive for a variety of reasons. First, it is difficult for listeners to analyze their discriminatory powers and describe quantitatively the important speaker discriminating features. Voiers [5] approached this problem by creating a list of 49 candidate differential factors, asking listeners to classify speakers according to these factors, then performing a factor analysis to determine the important speaker-discriminating features. Second, knowledge of these perceptual bases provides precious little insight toward the determination of productive acoustical features. In the Voiers study, the most significant speaker-discriminating features were "clarity," "roughness," "magnitude," and "animation," none of which can be related to acoustical parameters of the speech signal in a quantitative way. Finally, the direct use of these speaker-discriminating perceptual factors for speaker recognition is not very effective when compared with judgements made by actually listening to the speech data [10]. Thus the value of these specific perceptual features for speaker recognition is questionable.

Another approach to determining the potential for discrimination between speakers is to examine and statistically characterize the inventory of acoustical measurements of the speech signal as a function of speaker. One notable study of this sort attempted to assess discrimination potential as a function of phonetic class after first manually locating speech events within utterances [13]. Useful measures for discriminating among speakers were found to include voice-pitch frequency, the amplitude spectra of vowels and nasals, slope of the glottal source spectrum, word duration, and voice onset time. Of these, the voice pitch frequency exhibited the greatest speaker discrimination, as measured by the *F*-ratio of between- to within-speaker variance. Unfortunately, voice pitch is rather susceptible to change over time and it covaries strongly with factors such as speech effort level and emotional state.

III. Subjective Recognition of Speakers

Speaker recognition performance now becomes the major issue of this paper, including the level of performance that may be achieved under various experimental paradigms and conditions. First the subjective performance of humans is reviewed, followed by computer techniques. The subjective performance of humans usually relates to how reliably one may identify a person by listening to his voice. In addition, however, a great deal of interest has arisen during the past two decades regarding the use of visual identification of voice spectrograms for use in the courtroom.

A. Speaker Recognition by Listening

Speaker recognition by listeners has been studied broadly, typically with the motivation of learning something about how listeners recognize speakers. That is, studies have been made to investigate the variables that affect listener performance and to understand the perceptual bases of speaker recognition. In certain cases, however, it becomes important to know how *WELL* listeners can recognize speakers. In forensic applications particularly, it is important to be able to assess correctly the probative value of a listener's judgement regarding the identity of a speaker. It has frequently been stated that the judgement of listeners is not very reliable [9]. There is support, however, for the notion that listener performance compares favorably with that of other methods [8].

That performance is strongly influenced by acoustic and speaker conditions is intuitively obvious when considering the listening task. But definition of the listener's task probably has an even stronger impact on recognition performance for listeners than for machines. This is because the listening domain allows a wider selection of recognition environment and strategy. For example, listeners can recognize speakers without an explicit comparison of two speech samples, based upon acquaintance with the reference voices. This is, of course, the natural form of speaker recognition with which we are all familiar. Such decisions are likely to use a good deal of high-level idiosyncratic (including nonphysiological) information about the speaker, deriving from the extensive reference knowledge of the speaker. Thus listener performance in this task may be even better than in cases where the decision is based upon contemporaneous comparison with reference voices. (I know of no good comparative study, however.)

A recent interesting study of speaker recognition by listeners reports good performance without explicit reference speech tokens [40]. In this study, speech samples from 24 people were collected during the course of playing a

battleship game requiring communication between participants. Coworkers (in a work unit of about 40 people) were then tested on these samples to measure the listeners' speaker recognition performance. Measured performance was quite good, with an identification error rate of only 12 percent. This study is particularly interesting in that speaker recognizability was also measured using speech tokens processed through a narrow-band (2.4-kbit/s) LPC vocoder. In this case, the identification error rate rose to 31 percent.

A different task, related to the listening task stated above, is the task of listening to a set of reference samples and matching one of these samples to an unknown sample of speech heard at (and remembered from) some past time. This task is similar in form to the recognition of familiar speakers, except that in this case the unknown and reference samples are presented in reverse order (unknown first) and the speakers are not familiar to the listener. The reference samples are also typically limited to a short duration. This task is often a reasonable model for forensic applications in which the voice of an identity in question is not recorded, thus preventing detailed voice comparisons. It is intuitive that the performance of the listener will degrade with the time span between hearing the unknown sample and the references, and this has been borne out in one well-known study. In this study [1], listeners first listened to an unknown speaker read a paragraph, then at a later session they listened to five reference speakers (including the unknown) read the same paragraph. The speaker identification accuracy of the listeners in this experiment varied strongly as a function of the time interval between sessions, with accuracy of better than 80 percent correct for intervals less than one week dropping to chance performance for an interval of half a year.

We have seen that our understanding of the perceptual bases of speaker recognition does not support a useful model of our listening ability [5]. Perhaps a more rewarding study is the calibration of listener performance as a function of acoustic variables. Such variables include signal-to-noise ratio, speech bandwidth, amount of speech material used, and various speech transmission and coding systems. Such studies may also have a significance beyond the mere calibration of human performance. Specifically, if we assume that human listeners do very well (i.e., close to optimum) in speaker recognition tasks, then listener performance evaluations can give us good bounds on what is achievable. This information can be used, for example, to assess the potential for studies of machine recognition or to test the ultimate feasibility of practical applications.

Human listeners have proven themselves to be robust speaker recognizers when presented with degraded speech, with unflagging performance under significant spectral distortions, and noise conditions. For example, although it has been determined that the octave band of frequencies from 1 to 2 kHz is most useful to listeners [2], listeners have demonstrated only a modest doubling of speaker recognition error when the speech signal is severely high-passed at 2 kHz or low-passed at 1 kHz [7]. In this same study, the degrading effect of additive white noise was measured. Recognition performance remained high until the noise power exceeded the speech signal power. Although the performance of the listeners in this task domain was modest (20-percent error), the robustness of their performance under severe signal degradation should serve to inspire scientists striving to model the speaker recognition process.

More complex distortions of the speech signal have also been studied in a variety of contexts. In one study admitting of high performance (fixed text, contemporaneous comparison of unknown with reference), the effect on speaker verification performance of several speech coding systems (namely a 24-kbit/s ADPCM coder and an LPC pitch-excited vocoder) was measured [27]. In this study, several sentences were read by subjects and then processed through the ADPCM and LPC systems. Listeners, presented with two different tokens and asked to declare whether the speakers were the same or different, performed with relatively little degradation when the tokens were processed by different coding systems. As expected, the best performance was attained for the "no processing" control task (15 percent rejection of valid matches and 7 percent acceptance of invalid matches). But listener performance degraded surprisingly little when one of the two tokens was processed through either the ADPCM or LPC coding system. In this case, the likelihood of rejecting a correct match doubled while the likelihood of accepting an incorrect match increased very little.

Another very similar experiment has been conducted using, however, different texts for comparison [39] and evaluating the performance under degradations of telephone transmission and 2.4-kbit/s LPC coding. A speaker recognition task was evaluated, using five reference speakers and with the speakers reading phonetically balanced sentences. In all the tests, the speech of the reference speakers was presented in unprocessed form. Unprocessed tokens were misrecognized 14 percent of the time when they were taken from the same session as the reference speech and over 20 percent of the time when they were taken from a different session. Further, a significant increase in error rate was observed when the unknown tokens were processed through either the LPC system or the telephone (including transduction using a carbon button microphone), with the identification error rate for these two distortions being almost 50 percent.

Thus we see that, depending on the task definition and recognition conditions, listeners exhibit a wide range of speaker recognition performance. While it is clear that there are many sources of knowledge that are useful in the speaker recognition task, it is not clear which knowledge sources are most important for the achievement of high-recognition performance. It is likely that the various sources of knowledge contribute in varying ways to speaker recognition—providing weak, moderate, or high discrimination power and being more or less robust against various signal degradations and in various task definitions. More knowledge about these relationships would surely be helpful both in the development of automatic speaker recognition systems and in the assessment of speaker recognition by listeners in untested conditions.

B. Visual Identification of Speakers from Spectrograms

Probably the most significant paper on speaker recognition, as judged by the amount of further research it has stimulated, was a paper by Kersta introducing the spectrogram as a means of personal identification [3]. The term "voiceprint" was introduced in this paper, and 99-percent

correct identification performance based upon visual comparison of these voiceprints (spectrograms) was reported in a voiceprint identification task using 12 reference speakers.

The use of the term "voiceprint" has probably contributed to the popularity of voiceprint identification by analogy to the term "fingerprint." In fact Kersta himself, in his 1962 paper, disingenuously states: "Closely analogous to fingerprint identification, which uses the unique features found in people's fingerprints, voiceprint identification uses the unique features found in their utterances." Of course, as we have seen, the spectrogram is a function of the speech signal, not of the physical anatomy of the speaker, and it depends far more upon what the speaker does than upon what he "is." What the speaker does, in turn, is an indescribably complex function of many factors.

Unfortunately, the good performance reported in Kersta's paper has not been observed in subsequent evaluations simulating real-life conditions. In the largest evaluation of voiceprints ever conducted, under the direction of professor Oscar Tosi at Michigan State University, [14], 0.5-percent identification error was achieved using voiceprints for nine clue words under the restrictive conditions of isolated word utterances, closed trials, and contemporary speech. That is, the unknown speech tokens were spoken in isolation, the unknown speaker was known to be represented in the set of reference speakers, the unknown speech tokens were collected during the same data collection session as the reference tokens, and identification decisions were based upon comparison of spectrograms for all nine clue words. Unfortunately, these conditions do not fit any type of realistic voice identification scenario, particularly any type of forensic model, which is the major application of voiceprint identification. In an attempt to assess the performance of voiceprint identification in a more realistic environment, Tosi discovered that under conditions of nonisolated words, open trials, and noncontemporary speech the identification error rate escalated to 18 percent. Curiously, Tosi concludes that the voiceprint voice identification technique "could yield a negligible error" providing that more knowledgable examiners are selected and that they make decisions only when they are "absolutely certain" [16]. (Almost two thirds of the false identifications were judged as "uncertain.")

Speech scientists have tended to be critical of the voiceprint speaker recognition technique for a number of reasons [11]: the technique is not a well-defined objective procedure (art rather than science), the identification performance is strongly influenced by specific conditions which, without an underlying model, cannot be adequately forecast, and various evaluations of identification accuracy have been equivocal. A scientific committee sponsored by the National Research Council subsequently concluded that the experimental conditions covered by available evaluations do not constitute an adequate basis for making judgements of the reliability of voice identification in forensic applications [29]. For example, the speaker's emotional state, which can have dramatic impact upon the speech signal [17] and which is often an active element in the forensic model, has not been included in experimental conditions for controlled evaluations of voiceprint identification. In fact, there is evidence that the use of voiceprint identification has been extended far beyond its domain of usefulness in voice identification in actual criminal trials [20]. Thus without an objective means of calibration, the use of the voiceprint technique is dangerously susceptible to misuse.

The recognition reliability of voiceprints, relative to the reliability of a listener's judgement, is also an important consideration in the use of voiceprints (and in weighing voiceprint evidence in the courtroom). In previous studies comparing the performance of voiceprint identification with aural speaker discrimination by human listeners, the error rates for aural discrimination have always been smaller [6], [8], [10]. In the 1968 study by Stevens, for example, a closed-set identification test using a homogeneous group of eight reference speakers yielded 6-percent error for listening and 21-percent error for voiceprints. (In addition, the error rate for the voiceprint examiner with the best performance was still higher than that for the listener with the poorest listening performance.)

Thus the reliability of the voiceprint technique for speaker identification is clearly a fragile issue, because identification performance is sensitive to many acoustic, environmental, and speaker conditions. Furthermore, the use of the voiceprint technique is highly questionable, because better performance can likely be obtained through a listener's judgement. This brings up an important perspective on the development and evaluation of speaker recognition technology in general; namely, the comparative performance of a putative technique with respect to some generally accepted benchmark. Such a performance comparison seems to be a valuable step toward calibration of the absolute performance of any speaker recognition technique, be it a subjective one such as voiceprint examination or an objective one using computer-based speaker recognition algorithms.

IV. Computer Recognition of Speakers

Let us now turn our attention from speaker recognition by listeners to speaker recognition by computers. This technology has been an active research area for over twenty years, but with limited application success to date. Two excellent reviews of computer recognition of speakers were published in a previous special issue of these PROCEEDINGS [22], [23], and serve as an appropriate starting point for this review. In particular, Atal presents a general and reasonably comprehensive view of the selection of speech parameters for speaker recognition. This paper will discuss computer recognition of speakers in the context of two general application areas. These areas of interest are text-independent speaker recognition, in which the speaker is noncooperative, and text-dependent speaker recognition, in which the speaker is cooperative and interacts directly with the computer.

A. Text-Independent Recognition Technology

During the past few years, text-independent (or "free-text") speaker recognition has become an increasingly popular area of research, with a broad spectrum of potential applications. The free-text speaker recognition task definition is highly variable, from an acoustically clean and prescribed task description to environments where not only is the speech linguistically unconstrained but also the acoustic environment is extremely adverse. Possible applications include forensic use, automatic sorting and classification of

intelligence data, and passive security applications through monitoring of voice circuits. In general, applications for free-text speaker recognition have limited control of the conditions which influence system performance. Indeed, the definition of the task as "free-text" connotes a lack of complete control. (It may be assumed that a fixed text would be used if feasible, because better performance is possible if the text is known and calibrated beforehand.) This lack of control leads to corruption of the speech signal and consequently to degraded recognition performance. Corruption of the speech signal occurs in a number of ways, including distortions in the communication channel, additive acoustical noise, and probably most importantly through increased variability in the speech signal itself. (The speech signal may be expected to vary greatly under operational conditions in which the speaker may be absorbed in a task or involved in an emotionally charged situation.) Thus the free-text recognition task typically confers upon the researcher multiple problems—namely, that the input speech is unconstrained, that the speaker is uncooperative, and that the environmental parameters are uncontrolled.

Research into speaker characteristics and free-text recognition algorithms seems more appealing than fixed-text speaker recognition in a sense, because emphasis is on the search for features and characteristics unique to the individual rather than on artifactual differences that co-vary with particular phonetic environments. Nonetheless, performance of free-text speaker recognition has never approached that achievable within a controlled fixed-text task definition. Perhaps as a result of this, interest in and research on free-text recognition has historically lagged behind fixed-text work. During the last five years, however, research in free-text speaker recognition has matured greatly and interest in the free-text task is now quite high, judging by the relative amount of work in the area [33]–[35]. Focus has shifted from highly controlled databases and laboratory experiments to the processing of actual operational data. One consequence of this realism is that the level of recognition performance achieved lately has, unfortunately, deteriorated, with speaker recognition error rates not infrequently in excess of 20 percent [41].

One of the key issues in developing a text-independent speaker recognition system is to identify appropriate features and measures which will support good recognition performance. Use of the long-term average spectrum as a feature vector was discovered to have potential for free-text recognition during initial exploratory studies of fixed-text recognition using spectral pattern matching techniques [4]. In the Pruzansky study [4], speaker recognition error rate was found to remain undegraded (at 11 percent) even after averaging spectral amplitudes over all frames of speech data into a single reference spectral amplitude vector for each talker. To illustrate this feature vector, the long-term amplitude spectra for the different-speaker utterances shown in Fig. 2 are displayed in Fig. 4, and the long-term spectra for the same-speaker utterances of Fig. 3 are displayed in Fig. 5. Unfortunately, the long-term spectrum is not a good stable feature vector to use for speaker recognition. Long-term spectrum is obviously sensitive to changes in the spectral response of any interposed communications channel. More important, we have seen that the long-term spectrum is not particularly stable across variations in the speaker's speech effort level. A number of increasingly more sophisticated approaches have been developed to overcome some of the more fundamental limitations of a simple Euclidean distance measure on a simple spectral amplitude vector [26], [31], [34]. These approaches typically attempt to stabilize and statistically characterize features which represent the speech spectrum. These features include statistically orthogonal spectral vector combinations, cepstral coefficients, and a variety of LPC-based parameters. Surprisingly, the primary measure of choice remains the spectral amplitude vector, and very little effort has been devoted to the development of other measures such as pitch, formant frequencies, or statistical time functions. One reason for selecting the spectral amplitude vector is that it has typically produced performance superior to other features such as voice-pitch frequency [25].

Another key issue in free-text speaker recognition is the general strategy used to make a recognition decision. There have been two distinct approaches to this problem. First is the use of long-term averages. That is, certain features of the speech signal are computed for each incoming frame and are then averaged over a complete segment of speech. A recognition decision is made by computing the statistical

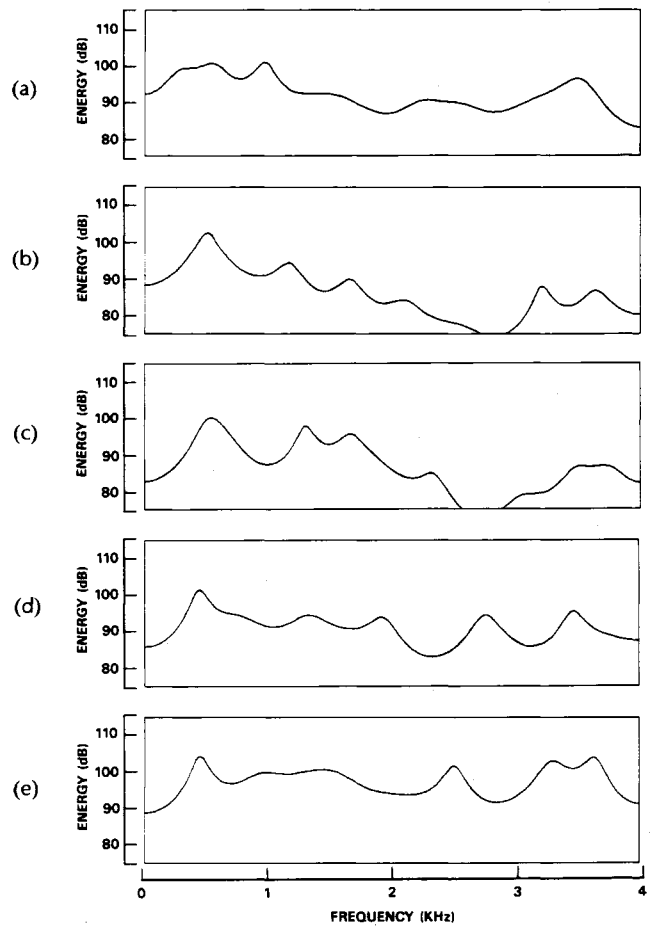

Fig. 4. This figure shows the long-term spectral amplitudes of the speech signals exhibited in Fig. 2, one for each of the five speakers. These spectra were computed over the extent of the utterance and were then smoothed by fitting the amplitude spectrum with a 20th-order LPC model. Notice that there exist significant differences between the spectra, despite the fact that the speakers all spoke the same utterance.

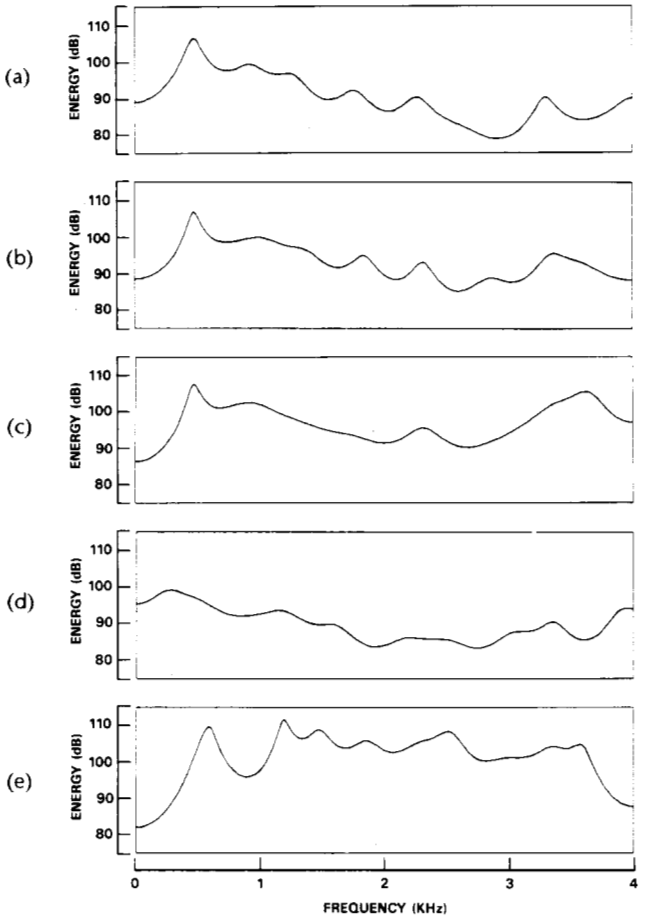

Fig. 5. This figure shows the long-term spectral amplitudes of the speech signals exhibited in Fig. 3, all for one speaker saying the same utterance under five different conditions. Notice that the spectra for (a) and (b), which were collected at different sessions but otherwise under the same conditions, are quite similar, except perhaps above 3 kHz. The other long-term spectra demonstrate large differences attributable to the various conditions of (c) carbon button microphone, (d) soft (−20-dB) speech, and (e) loud (+20 dB) speech.

likelihood of the long-term average feature vector given the various speaker hypotheses. Most work has used this approach, and this approach appears to yield the best performance today and is therefore most popular [31], [34]. The second approach, which is intuitively appealing, is to search for specific phonetic events in the incoming speech signal and then to compare the speech features of the selected and detected phonetic events with those features belonging to the matching phonetic event of the reference speakers. The problem with this approach is that errors in detecting phonetic events tend to corrupt the speaker recognition process. Even using speaker-specific phonetic references, the reliability of phonetic detection is currently inadequate to support good speaker recognition performance [28]. A modification of this approach which has provided reasonably good recognition performance is to place a tight threshold on the detection of these reference phonetic events and then to determine the speaker based upon the frequency of detecting a speaker's phonetic events [21], [35]. Another more recent recognition approach avoids the problems of phonetic detection while benefiting from short-term phonetic information about the speech signal [33]. In this approach, the short-term feature vectors are characterized statistically for the reference speakers, then the likelihood of the input speech is computed based upon this statistical model. Such an approach has been applied to difficult speaker recognition problems involving noise and distortion with some success [38].

A critically important aspect of the development of a free-text speaker recognition system is the database used for its development and evaluation. In the earliest studies, subjects read from prepared texts in an environment relatively free of noise. This caused a degree of skepticism about the significance of such results in an operational environment. An important study which addressed this problem was performed by Markel and Davis [30]. In this effort, a linguistically unconstrained database of extemporaneous speech was collected from eleven men and six women over a period of three months. This database was then used to develop a highly successful speaker recognition technology. The features used in this system were based upon the voice pitch and the reflection coefficients of an LPC-10 model. The recognition features were the mean and the standard deviation of these eleven parameters. Using a statistically orthogonal linear transformation of these features, performance of 2-percent identification error (and 4-percent verification error) was achieved for 40-s segments of input speech. This result is excellent for unconstrained extemporaneous speech. Of course, poorer results were obtained for shorter speech segment durations, and recognition error exceeded 20 percent for segment durations below 4 s.

Although the Markel database stands out as a significant step forward in simulation of realistic applications, it does not demonstrate many of the difficulties of operational data. For example, microphone degradation, acoustic noise, and channel distortion were not addressed. A recent effort at Bolt Beranek and Newman Inc. has calibrated the difficulty of free-text speaker recognition on an operational database collected over a radio channel [37], [38]. This database is corrupted by noise (with an average S/N ratio of 19 dB) and emotionally stressful task-oriented activity. Further, the segment durations used for recognition are short. On this database, the best performance achieved over a variety of transformations of the spectrum yielded 30-percent error for segment durations of 2 s on a population of nine male speakers. When cross-channel recognition was attempted (enrolling on data from one radio channel, then attempting recognition on another channel) the error rate rose further to about 50 percent. Attempts at channel equalization were effective only to a modest degree, with up to 5-percent improvement in recognition error rate. This is a sobering demonstration of the difficulty of the operational recognition task.

Performance is determined in speech tasks by the quality of the speech database evaluated, and excellent performance is often quite easy to achieve if the speech data are carefully controlled. As we have seen, recognition error can vary by an order of magnitude, depending on the difficulty of the database. Because of this, it is impossible to make serious comparisons of different recognition approaches unless they are evaluated on the same database. There are several possible solutions to this difficulty. One is simply to use a single standard database for evaluation of all recogni-

tion techniques. This will not often be satisfactory because the diversity of applications will demand consideration of many divergent database conditions. A more viable solution would be to calibrate algorithm performance against human listener performance on the same task. Human listeners are generally regarded to be good speaker recognizers who maintain relatively good performance under most degrading conditions. This calibration technique, therefore, offers a means of benchmarking the performance for any speaker recognition algorithm and a means of comparing the performance of systems in different application scenarios. Such a technique has been useful in a variety of applications [18], [39].

B. Fixed-Text Verification Technology

Of all the forms of automatic speaker recognition technology, the one with the greatest potential for practical application is fixed-text speaker verification. Speaker verification has the potential to add security and convenience to home door locks, automobile ignition switches, automatic teller machines, and bank-by-phone facilities. Fixed-text speaker verification is the form of speaker recognition used for security applications in which a person desires privileged entry or access to some protected resource. This privilege is granted upon verification of the person's identity, which is performed by comparing his voice characteristics with that of a valid user. The speaker in these applications is thus cooperative, which helps immeasurably in achieving good speaker recognition performance. First, he is willing to proffer his identity to the system, which reduces the identification process (who is he?) to a verification process (is he truly who he claims to be?). Second, he is willing to say whatever is requested of him. Finally, he is willing (although perhaps not completely able) to say the requested speech token consistently during each verification attempt.

Because of the high degree of control which can usually be exercised over the speech signal conditions, performance in fixed-text verification is typically much better than for other speaker recognition tasks where the degree of control over the recognition environment is limited. In fact, performance of fixed-text verification has in many cases reached the point where practical application of the technology is being considered [23]. One of these applications, to control physical entry, will be described in detail in the next section.

Perhaps the most compelling applications of speaker verification, however, involve the verification of voices transmitted over telephone lines, where corruption of the speech data by microphone and channel characteristics remains a difficult problem. Although limited success has been achieved, speaker verification over the telephone still offers a challenge to those interested in the development of practical voice verification technology. One notable study of voice verification over the telephone examined the performance of the Bell Laboratories voice verification system for more than 100 men and women in a realistic operational simulation over an extended period of 5 months [24]. This system, which used as speaker discrimination features only the speech energy and voice pitch as a function of time, exhibited a user rejection rate and impostor acceptance rate of about 10 percent initially for new users, with error rate declining to about half this value for experienced users and fully adapted speaker templates. Another interesting result of this study was the histogram of population statistics versus performance, with about half of the users of the system experiencing less than 5-percent error while a small but significant fraction of the population exhibiting over 20-percent user rejection or impostor acceptance.

The most typical strategy for performing fixed-text speaker verification is to create a reference file of speech parameters (as a function of time) for each user and then, during verification, to compare the speech parameters for the unknown speaker with reference parameters at equivalent points in time. That is, the input speech from the unknown speaker is time aligned with the reference speech for the proffered identity, and the distance between corresponding times is computed and averaged over the utterance. This general technique has served as a basis for almost all approaches to fixed-text speaker verification. An exception to this general strategy is the approach that computes time-averaged statistics of speech frame parameters and bases the verification decision on the estimated likelihood of the reference speaker given these statistics. This latter approach is very much like that for free-text verification, except that the verification utterance is fixed and therefore the time-averaged statistics are far more stable than if the speaker's utterances were unspecified.

These two approaches, namely, comparing dynamic speech features at equivalent points in time and comparing average speech feature statistics, have been compared and found to yield similar performance in at least one study [32]. In the Furui study, the speech features included the log area ratios of a 12th-order LPC model and the voice pitch frequency, and good verification performance was achieved for a set of nine men speaking two short Japanese words using either the time-averaged statistical features or the dynamic features. Less than 1-percent error was achieved on input utterances spoken 10 months after enrollment. Error rate increased to more than 3 percent on utterances spoken 5 years after enrollment, however, thus indicating the desirability of periodically updating the speakers' reference data. This good performance was also critically dependent upon careful equalization of the average speech spectrum. Spectral equalization was performed by filtering the input signal with a two-zero critically damped filter adjusted so as to flatten the average input spectrum. This spectral equalization was shown by Furui to reduce the error rate by as much as a factor of two.

A concern in the operational use of speaker verification is the susceptibility of the verification algorithms to mimicking. Unfortunately, few formal studies have been conducted on this issue. In one study performed at Bell Laboratories [15], four professional mimics were selected from a much larger set of candidates. These four, who sounded best in the prescreening trials, were then coached intensively and attempted mimicking enrolled users under favorable conditions. The mimics did a rather good job on timing and inflection and achieved an order of magnitude greater impostor acceptance rate (27 percent on the automatic system) than did casual impostors. Later versions of this system, which stressed spectral features rather than prosodic features, reduced this mimic error rate considerably, however. An interesting note on mimicry is the comparison of human versus machine performance on dis-

tinguishing the voices of identical twins. In the Bell Labs database there was a pair of identical twins, and listeners who otherwise did well in discriminating speakers almost universally accepted the twin as his brother (96-percent acceptance) [18]. Interestingly, the machine never confused the two twins. This suggests that man and machine are using different features or strategies in making verification decisions, despite the fact that their overall performances are comparable.

There are limits to the performance achievable with voice verification, as previously discussed, and these limits in the speaker verification task often relate to the degree of voice consistency that the user can maintain. This is, among other things, a worthy human factors challenge for the system designer. How do you ensure that the user uses the same rate of speech and the same speech effort level for each verification? More importantly, how do you prevent the users from contracting ailments such as the common cold that affect the voice quality? The essence of the challenge is to control the rare statistics rather than the mean. "Typical" speech input may always yield perfect performance, so that the error performance of a system may be determined by the frequency of occurrence of anomalous speech data. Shrewd control of the system users, along with the use of robust speech features, are key factors in establishing a high-performance speaker verification system.

V. An Operational Speaker Verification System

Texas Instruments currently uses a voice verification system to control physical entry into its main computer center at corporate headquarters in Dallas. A brief description of this system will serve to illustrate some of the human factors problems encountered and solutions developed to make a successful operational system. This system has been operational 24 h per day for more than a decade now. To use the system an entrant first opens the door to the entry booth and walks in, then he identifies himself by entering a user ID into a keypad, and then he repeats the verification phrase(s) that the system prompts him to say. If he is verified, the system says "verified, thank you" and unlocks the inside door of the booth so that he may enter into the computer center. If he is not verified, the system notifies him by saying "not verified, call for assistance."

The verification algorithm is fairly straightforward, using as a speech feature vector the output of a 14-channel filter bank whose frequencies are spaced uniformly between 300 and 3000 Hz. The verification decision is based upon the cumulative Euclidean distance between the features of the speaker's reference frames and those of the time-aligned input frames. Time alignment is established at the point of best match between input and reference feature vectors using a simplified form of dynamic time warping.

Verification utterances are constructed randomly to avoid the possibility of being able to defeat the system with a tape recording of a valid user. An innocuous four-word fixed phrase structure is used, with one of sixteen word alternatives filling each of the four word positions. The complete set of words is shown in the Table 2. An example verification utterance might be "Proud Ben served hard." These utterances are prompted by voice. This is thought to improve verification performance by stabilizing the pronunciation of the user's utterance. (The user will tend to say the utterance in the same way as it is prompted. We are concerned about maintaining highly consistent and reliable user speech input and relatively unconcerned about possible help to the impostor through this voice prompt template.)

Table 2 Verification Phrase Construction for the TI Operational Voice Verification System

GOOD	BEN	SWAM	NEAR
PROUD	BRUCE	CALLED	HARD
STRONG	JEAN	SERVED	HIGH
YOUNG	JOYCE	CAME	NORTH

Each reference word is represented by a template of six frames in which the reference frames are spaced 20 ms apart. Each of the 14 normalized filter amplitudes is represented as a 3-bit number, but the reference data are stored as 8-bit numbers, with 5 fractional bits allocated to accommodate adaptive updating of the reference templates. Thus the basic reference speech data storage requirement is 1350 bytes per speaker. The processing requirements of this system are also reasonably modest. During verification, the processing is dominated by Euclidean distance computations. Each input frame must be compared with every active reference frame, and with an input frame period of 10 ms (the input frame period is half of the reference frame period) the basic multiply-accumulate rate is 34 000 multiply-accumulates per second.

A single utterance is inadequate to provide the high level of verification performance desired. Therefore, a sequential decision using multiple verification phrases is employed which provides less than 1-percent rejection of users and less than 1-percent acceptance of impostors. Four phrases are constructed randomly at the outset of a verification attempt so that all sixteen reference words are used, and all four phrases may be used if required during a verification attempt. In fact, phrases that have been used may be reprompted and reused, which is sometimes necessary because of occasional mispronunciations. Thus up to seven user utterances may be solicited in the course of a verification attempt, although the number as measured in operational usage averages only 1.6.

The gross rejection rate of the operational system has been measured as 0.9 percent, with a casual impostor acceptance rate of 0.7 percent. The term "gross rejection rate" is used rather than user rejection rate, because the system is clearly justified in rejecting some of the entrants. For example, a special rule states that if the entrant makes no response for two consecutive prompts, then the verification is aborted. Twenty percent of all rejections (0.2 percent of all entry attempts) are attributable to this special rule. Another interesting correlation with rejection rate is the number of people in the entry booth. Since the floor of the booth is a weight scale, it is possible to accomodate multiple entrants. This is achieved by recording the weight of each user during his enrollment and then by counting the number of people in the booth during verification. It is curious to note that the rejection rate is much lower with only one person in the booth (0.5 percent) than it is with more than one person in the booth (1.8 percent). A similarly

curious correlation exists between rejection rate and time of day, with the rejection rate between 9 am and 3 pm being four times lower than the rejection rate between 9 pm and 3 am. Indeed, recordings of booth activity for multiple entrants during the night shift are at times rather entertaining!

Another important factor in operational speaker verification is enrollment. Specifically, enrollment is a problem because the initial estimate of a person's speech characteristic taken at a single session is likely to be biased by the speaker's momentary speech qualities, so that a relatively poor match is sometimes encountered during the first few sessions after enrollment. Perhaps an even more troublesome problem is the dramatic change that often occurs in a user's voice during the first few entry attempts. During enrollment the typical new user is intimidated by the system, and this intimidation can easily affect the user's voice, for example in changes in the loudness, rate of speech, and pitch frequency. Although the reference data are updated by averaging with the input data after each successful verification, rapid changes such as during this initial learning period and also at the onset of respiratory ailments may result in serious rejection problems. An example of this is demonstrated by observing the user rejection rate as a function of user experience. Specifically, during the first four verification attempts the user rejection rate has been measured as just under 10 percent! Immediately after this four-session orientation period, however, the rejection rate drops to 1 percent and then remains uniformly less than 1 percent. Out to about 1000 verification trials the rejection rate hovers at slightly below 1 percent, then it gradually declines to less than one quarter of a percent for users with more than 10000 verifications.

Another difficulty in achieving satisfactory total system performance is the inhomogeneity in performance across a population of users. Specifically, it is generally observed that verification performance is very good for most users, but that for a small percentage of users the performance is not so good. This has been observed for the TI operational system by measuring the distribution of rejection rate as a function of user. This distribution reveals that a typical user (i.e., having the median rejection rate) exhibits a rejection rate of only one half the average rejection rate. This suggests that most of the rejections are clustered in a small portion of the population (which we refer to as "goats," in contrast with the better performing "sheep"). Indeed, it is observed that only one quarter of the population exhibit rejection rates greater than the average rejection rate.

VI. Speaker Recognition in the Future

Fifteen yeas ago, when I first became involved in speech technology, I was frankly not very optimistic about the prospects for commercial application of automatic speech recognition and speaker recognition technology. This was not so much because the technology was deficient, but more because economical solutions were beyond imagination. Now, of course, computers weigh grams rather than tons, and the cost has gone down almost proportionally. In fact, the cost problem has been virtually eliminated. We have finally reached the point, with the advent of high-speed monolithic digital signal processor circuits, where the cost of the speech processing, per se, is potentially a small fraction of the cost of the complete system. Yet we still see very little commercial activity in the area of speaker recognition. Why is this?

The most obvious application of speaker recognition technology is in security systems, for physical entry control, or for various business transactions conducted over the telephone. But the product development activities in these areas actually seems to have subsided during the past five years. There are a number of possible factors which have contributed to this decline. First, this lack of business interest in speaker recognition is surely partially attributable to remaining inadequacies in speaker recognition performance. As in speech recognition, machine performance for speaker recognition tends to be rather "fragile," that is, sensitive to distortions and variations in the speech signal that are innocuous or even imperceptible to human listeners. This lack of "robustness" results in errors, and worse, these errors are difficult for the user to "understand" and may therefore leave the user indignant with an exaggerated perception of the offense.

Other factors not directly related to recognition performance may play even more important roles in hindering the adoption of speaker recognition technology. Most of these factors are system complexity factors which are almost a natural consequence of the use of a high-performance man/machine interface. Such factors have also limited the growth of speech recognition products. For example, the interaction between the system and the user must be given special consideration in all phases of system development, which causes major impact on overall system design. (The speech subsystem is not just something that is neatly grafted onto an otherwise self-contained system.) This interdependence between the speech technology and the application host extends to many major design decisions, such as:

*Where should the reference speech data be stored? Centrally at the application host, or peripherally at the terminal, or perhaps with the user as a magnetic stripe card. (Speaker recognition typically requires a large amount of reference data per user, perhaps as much as 1 kbyte/user or more for a high-performance system.) The handling of this problem bears on the cost, the security, and usage protocol of the system.

*Where will the users be enrolled?

*How will information be communicated between the security terminal and the central host? (Audio lines may be preferable for short links, but terminal-based digital signal processing may be mandatory for long links where line degradation might be prohibitive.)

*How will the system handle user rejections? (This is a question that must be answered in operational systems, in such a way that the security of the system is maintained while the cost and feasibility are kept under control.)

Finally, perhaps the most important factor which is inhibiting the widespread adoption of speaker recognition technology for security applications is the lack of a truly compelling need. Although speaker recognition is an appealing technology for enhancing security in physical entry control and data access control applications, the benefits afforded by voice apparently have not yet outweighed the burden and expense of installing such a system.

It may be that voice verification will never catch on as a premium method of identity validation. However, there are indications, as our society edges forward into the abyss of total automation and computerization, that the need for

personal identity validation will become more acute. This need may be even more important when considered in the context of consumer business telecommunications. Specifically, the time, effort, and cost required for physical transportation to conduct personal business becomes more prohibitive in proportion to the richness of our lives and our desire to enjoy every minute of it. So there has arisen a vast array of services that are now available over the telephone —airline reservations, bank-by-phone, and telephone-order catalog services. These automatic telephone transaction systems could likely be facilitated if a more secure means of personal authentication were available. And of course, over the telephone, the use of the voice signal seems an ideal choice.

In the course of speaker recognition technology development it is important to bear in mind the nature of the problem. Specifically, it is not reasonable to expect that any level of performance is possible, limited only by improvements in feature extraction and algorithm development. Rather, it should be clear from the preceding discussion and illustrations in this paper that the performance of a speaker recognition system is dependent upon the amount of control that can be exerted on the operational conditions. And of greatest importance among all the variables is the speaker. For well-controlled conditions with a cooperative user operationally satisfactory performance might be obtained in a security system. On the other hand, for uncontrolled conditions and with uncooperative speakers, it may be impossible to assure any particular level of performance.

The fact that speaker recognition performance achievable on any particular task is a sensitive function of the data as well as the task definition strongly suggests the need for benchmark databases. The use of benchmark databases for speech recognition evaluations has grown in popularity during the last five years, in response to the need for meaningful comparative evaluation of systems and in view of the extreme performance sensitivity to particular databases. The need for such benchmarking databases for speaker recognition technology is acute for research into automatic recognition technology as well as for evaluation of proposed systems. This need is particularly acute for forensic applications of speaker identification. In this difficult application environment it is critically important to assess and compare the performance of the various techniques that are currently being used and that are being proposed. How do lay human listener judgements compare with voiceprint examination? Can speaker recognition performance be enhanced by combining listening and visual examination of spectrograms? What are reasonable bounds on speaker verification performance as a function of realistic task conditions? These questions remain largely unanswered.

Performance evaluation is also a difficult issue for security applications. These systems are invariably text-dependent systems, and the text definition is often considered to be part of the system definition. This makes comparative evaluation of these systems difficult. The database definition for free-text systems, on the other hand, is somewhat easier since the text is unconstrained and cannot be specified by the application. Of course, there exist endless varieties of degrading conditions that might be defined as part of an evaluation database, but it appears quite feasible and desirable to establish a benchmark database for the evaluation of free-text speaker recognition technology. In the definition of these evaluation databases emphasis really should be placed upon enriching the database by including as many difficult conditions into the database as possible. This provides two key benefits. First, technology development is facilitated by focusing attention and research in the important problem areas. Second, research efficiency is increased by attaining a given level of statistical confidence in the performance of the system (i.e., number of errors) with a significantly smaller database and correspondingly less computational and storage burden.

Finally, I would recommend that research be directed more seriously toward the problem of achieving high-performance speaker verification over standard telephone circuits. This is a difficult problem because of the variable carbon button microphone transducer and transmission channel, but a satisfactory solution to this problem would facilitate telephone business services and could help to revolutionize the way people do their personal business. As a first step, it might be advisable to calibrate the performance of human listeners carefully and determine if an adequate level of listener verification performance is feasible by appropriate choice of speech material. Then attempt to achieve this level of performance by machine. What is an "adequate level" of performance? That is a most difficult question and one that can probably only be determined by a complete system definition and perhaps then by trial and error.

References

[1] F. McGehee, "The reliability of the identification of the human voice," *J. General Psychol.*, vol. 17, pp. 249–271, 1937.

[2] R. W. Peters, "Studies in extra messages: The effects of various modifications of the voice signal upon the ability of listeners to identify speakers' voices," NM 001-104-500, Joint Report 61, Pensacola, FL, USNSAM, 1954.

[3] L. G. Kersta, "Voiceprint identification," *Nature*, vol. 196, pp. 1253–1257, 1962.

[4] S. Pruzansky, "Pattern-matching procedure for automatic talker recognition," *J. Acoust. Soc. Amer.*, vol. 35, pp. 354–358, 1963.

[5] W. D. Voiers, "Perceptual bases of speaker identity," *J. Acoust. Soc. Amer.*, vol. 36, no. 6, pp. 1065–1073, 1964.

[6] J. R. Carbonell *et al.* "Speaker authentication techniques," BBN report 1296, Cambridge, MA, Bolt Beranek and Newman Inc., 1965.

[7] F. R. Clarke *et al.*, "Characteristics that determine speaker recognition," Electronic Systems Division, USAF, Tech. Rep. ESD-TR-66-636, 1966.

[8] K. N. Stevens *et al.*, "Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material," *J. Acoust. Soc. Amer.*, vol. 44, pp. 1596–1607, 1968.

[9] R. H. Bolt *et al.*, "Identification of a speaker by speech spectrograms," *Science*, vol. 166, pp. 338–343, 1969.

[10] F. R. Clarke and R. W. Becker, "Comparison of techniques for discriminating among talkers," *J. Speech Hearing Res.*, vol. 12, pp. 747–761, 1969.

[11] R. H. Bolt *et al.*, "Speaker identification by speech spectrograms: A scientist's view of its reliability for legal purposes," *J. Acoust. Soc. Amer.*, vol. 47, no. 2, pp. 597–612, 1970.

[12] M. Hecker, "Speaker recognition: An interpretive survey of the literature," ASHA Monograph 16 (Amer. Speech and Hearing Assoc.), Washington, DC, 1971.

[13] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J. Acoust. Soc. Amer.*, vol. 51, no. 6, pp. 2044–2056, 1972.

[14] Anon., "vice identification research," Law Enforcement Assistance Administration, U.S. Department of Justice, Rep. PR 72-1, no. 2700-0144, 1972.
[15] R. C. Lummis and A. E. Rosenburg, "Test of an automatic speaker verification method with intensively trained professional mimics," *J. Acoust. Soc. Amer.*, vol. 51, no. 1 pt. 1 (abstract), 1972.
[16] O. Tosi *et al.*, "Latest developments in voice identification," *J. Acoust. Soc. Amer.*, vol. 51, no. 1 (abstract), 1972.
[17] C. E. Williams and K. N. Stevens, "Emotions and speech: Some acoustical correlates," *J. Acoust. Soc. Amer.*, vol. 52, pp. 1238–1250, 1972.
[18] A. E. Rosenburg, "Listener performance in speaker verification tasks," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, no. 3, pp. 221–225, 1973.
[19] G. R. Doddington, "Speaker verification," Final Rep. RADC-TR-74-179, Rome Air Development Center, Griffiss Air Force Base, Rome, NY, 1974.
[20] F. Poza, "Voiceprint identification: Its forensic application," in *Proc. Carnahan Conf. on Crime Countermeasures* (Lexington, KY, 1974).
[21] G. R. Doddington and M. Hydrick, "Speaker verification II," Final Rep. RADC-TR-75-274, Rome Air Development Center, Griffiss Air Force Base, Rome, NY, 1975.
[22] B. S. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 6, no. 4, pp. 460–475, 1976.
[23] A. E. Rosenberg, "Automatic speaker verification: A review," *Proc. IEEE*, vol. 64, no. 4, pp. 475–487, 1976.
[24] ———, "Evaluation of an automatic speaker verification system over telephone lines," *Bell Syst. Tech. J.*, vol. 55, no. 6, pp. 723–744, 1976.
[25] J. D. Markel, B. Oshika, and A. H. Gray, Jr., "Long-term feature averaging for speaker recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, no. 4, pp. 330–337, 1977.
[26] R. S. Cheung, and B. Eisenstein, "Feature selection via dynamic programming for text-independent speaker identification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 397–403, 1978.
[27] C. A. McGonegal, L. R. Rabiner, and B. J. McDermott, "Speaker verification by human listeners over several speech transmission systems," *Bell Syst. Tech. J.*, vol. 57, no. 8, pp. 2887–2900, 1978.
[28] L. Pfeifer, "New techniques for text-independent speaker identification," in *Proc. ICASSP-78*, pp. 283–286, 1978.
[29] R. H. Bolt *et al.*, "On the theory and practice of voice identification," National Academy of Sciences, Washington, DC, Rep. 0-309-02873-16, 1979.
[30] J. D. Markel and B. Davis, "Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 1, pp. 74–82, 1979.
[31] R. E. Wohlford, "A comparison of four techniques for automtic speaker recognition," in *Proc. ICASSP-80*, pp. 908–911, 1980.
[32] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, no. 3, pp. 342–350, 1981.
[33] R. Schwartz, S. Roucos, and M. Berouti, "The application of probability density estimation of text-independent speaker identification," in *Proc. ICASSP-82*, pp. 1649–1652, 1982.
[34] M. Shridhar and N. Mohankrishnan, "Text-independent speaker recognition: A review and some new results," *Speech Commun.*, vol. 1, pp. 257–267, 1982.
[35] K. P. Li and E. H. Wrench, Jr., "An approach to text-independent speaker recognition with short utterances," in *Proc. ICASSP-83*, pp. 555–558, 1983.
[36] F. Nolan, *The Phonetic Bases of Speaker Recognition.* Cambridge, UK: Cambridge Univ. Press, 1983.
[37] J. Wolf and M. Krasner, "Further investigation of probabilistic methods for text-independent speaker identification," in *Proc. ICASSP-83*, pp. 551–554, 1983.
[38] M. Krasner *et al.*, "Investigation of text-independent speaker identification techniques under conditions of variable data," in *Proc. ICASSP-84*, paper 18B.5, 1984.
[39] P. E. Papamichalis and G. R. Doddington, "A speaker recognizability test," in *Proc. ICASSP-84*, paper, 18B.6, 1984.
[40] A. Schmidt-Nielen and R. Stern, "Identification of known voices as a function of familiarity and narrow-band coding," *J. Acoust. Soc. Amer.*, vol. 77, no. 2, pp. 658–663, 1985.
[41] H. Gish *et al.*, "Investigation of text-independent speaker identification over telephone channels," in *Proc. ICASSP-85*, paper 11.1.1, 1985.

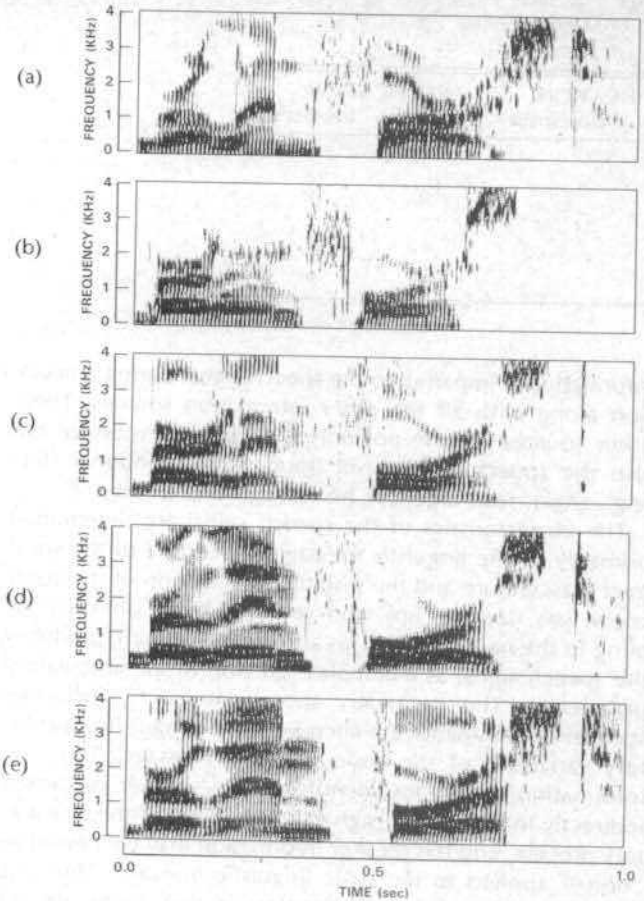

Fig. 2. This figure exhibits five different spectrograms, one each from five different men. The spectrogram is a display of the amplitude of a speech signal as a function of frequency and time. The spectral amplitude is computed as a short-time Fourier transform of the speech signal and is plotted on the *z*-axis, with greater spectral amplitude being depicted as a darker marking. The running window used in the spectral analysis is 6 ms long and the signal is weighted by a Hamming window. The abscissa is the time axis, with 1 s being displayed in this figure. The ordinate is the frequency axis, which spans the range from 0 Hz to 4 kHz in this figure. Although the spectrogram represents amplitude only imprecisely, a great deal of phonetic information may be decoded from the spectrogram by an expert phonetician. Indeed, many studies have shown that the energy loci in frequency and time are perceptually more important than exact calibration of the spectral amplitudes. (These energy loci are usually referred to as "formant" frequencies by phoneticians.) The utterance spoken in each of the five speech spectrograms displayed in this figure is "Berlin Forest." Note that although there are general similarities in the spectrograms as dictated by the linguistic message, the speaker differences are striking. So striking, in fact, that one might be tempted to question the sameness of the phonetic transcription. Two of the speaker represented in this figure are identical twins, and their voices do sound quite similar. These are speakers (b) and (c). Even for these twins, the spectrograms are unquestionably different.

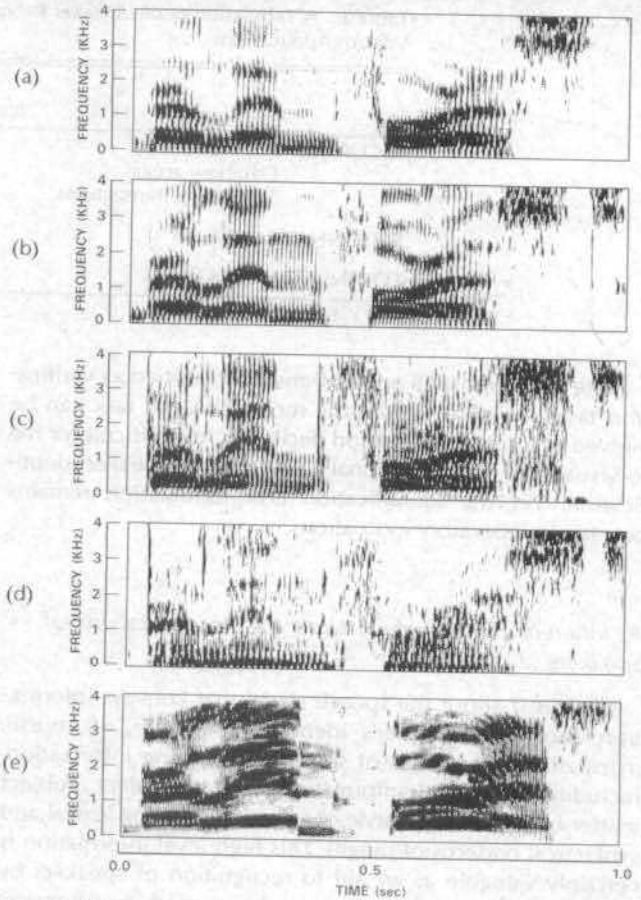

Fig. 3. This figure exhibits five different speech spectrograms, all from the same speaker, one who is not represented in Fig. 2. The utterance spoken in these spectrograms is the same as in Fig. 2. Note that the variation in spectrograms produced by the same person can be extremely great, even greater than the differences between speakers seen in Fig. 2. Examples (a) and (b) are for speech under nominal conditions, but taken from two different recording sessions. There are some significant amplitude differences, particularly above 2 kHz, but the formant frequencies remain nearly identical in frequency and time. Example (c) is for speech collected through a carbon button microphone. Notice that the nonlinearities of this microphone create significant spectral distortions and that the weaker formant frequencies above 21 kHz are largely obscured. Example (d) is for very softly spoken speech (about 20 dB below a normal comfortable level). Notice several changes; First, the spectrum falls off more rapidly with frequency, with most of the signal energy appearing below 1 kHz. Second, the spectrum appears "noisy," which is largely attributable to irregular voiced excitation. These changes also make the formant frequencies much less distinct. Example (e) is for very loudly spoken speech (about 20 dB above a normal comfortable level). The spectrogram of this speech signal bears little resemblance to that in (a) (at least relative to the expected between-speaker differences exhibited in Fig. 2), with a much higher pitch frequency and with relatively more energy at the higher frequencies. Indeed, the two speech signals do not sound much like the same person, either.