PAMUKKALE ÜNİVERSİTESİ MÜHENDİSLİK FAKÜLTESİ
PAMUKKALE UNIVERSITY ENGINEERING COLLEGE
MÜHENDİSLİK BİLİMLERİ DERGİSİ
JOURNAL OF ENGINEERING SCIENCES
YIL : 2000
CİLT : 6
SAYI : 2-3
SAYFA : 185-193

FUNDAMENTALS OF SPEAKER RECOGNITION

Figen ERTAŞ
Uludağ University, Faculty of Engineering and Architechture, Electronic Engineering Deparment, Görükle/Bursa

Geliş Tarihi : 11.10.1999

ABSTRACT

The explosive growth of information technology in the last decade has made a considerable impact on the design and construction of systems for human-machine communication, which is becoming increasingly important in many aspects of life. Amongst other speech processing tasks, a great deal of attention has been devoted to developing procedures that identify people from their voices, and the design and construction of speaker recognition systems has been a fascinating enterprise pursued over many decades. This paper introduces speaker recognition in general and discusses its relevant parameters in relation to system performance.

Key Words : Voice, Speaker, Recognition, Verification, Identification

KONUŞMACIYI TANIMANIN ESASLARI

ÖZET

Bilişim teknolojilerinde son yıllarda meydana gelen hızlı gelişmeler, yaşamın bir çok alanlarında önemi gittikçe artan insan-makine iletişim sistemlerinin tasarım ve gerçeklenmelerini önemli ölçüde etkilemiştir. Çeşitli söz işleme uygulamaları arasında, kişileri seslerinden tanıma yöntemleri geliştirmek ayrı bir önem kazanmış ve konuşmacıyı tanıyan sistemlerin tasarımı ve gerçeklenmeleri uzun yıllar üzerinde durulan cazip bir yatırım alanı olmuştur. Bu makale, ilgili performans parametrelerini tartışıp, konuşmacı tanımayı genel olarak tanıtmaktadır.

Anahtar Kelimeler : Ses, Konuşmacı, Tanıma, Doğrulama, Belirleme

1. INTRODUCTION

One of the most interesting and exciting areas of speech communication is man-machine communication by voice. Being able to communicate to computers, and have them understand what is said and also recognise who is speaking, would provide a comfortable and natural form of communication. The design and construction of speech recognition and speaker recognition systems has been a fascinating enterprise pursued over many decades. This paper introduces speaker recognition in general and discusses its relevant parameters in relation to system performance. Speech parameters with their relevance and suitability for speaker recognition purposes are also discussed.

2. SPEAKER RECOGNITION

Having computer procedures that understand spoken messages, there is also considerable interest in developing procedures that identify people by means of measurements on their voice signals. The ability to recognise a person from his voice is known as speaker recognition and this has recently received a great deal of attention among speech researchers. Since the performance of the human in discriminating among speakers has long been known, the most important aim in this field is to find out if computers could be programmed to recognise speakers from their voices as well as humans can. In many speech applications, it is difficult to say whether duplicating human performance by machines is manageable, but in speaker recognition, this is not true. Current experimental evidence in the literature indicates that

machines, particularly in speaker verification tasks, with short utterances and large number of speakers, have the potential to perform better than human listeners. This is particularly true for unfamiliar speakers, in which the "training time" for humans to learn a new voice well is very long compared to that for machines (Atal, 1976; Jesorsky, 1978; O'Shaughnessy, 1986; Bennani and Gallinari, 1994; Lastrucci et all., 1994).

Personal identification and/or verification is an essential requirement for controlling access to protected resources. Personal identity can be claimed by a key, a password, or a badge, all of which can be easily stolen, lost, faked or disguised. However, there are some unique (biometrics) features of individuals which cannot be imitated by someone else. Biometrics uses physical characteristics such as fingerprints, hand geometry and retinal pattern, and personal traits such as handwriting and voiceprint (Woodward, 1997). Although fingerprints or retinal patterns are usually more reliable ways of verifying that a person is who he claims to be, identity verification based on a person's voice has special advantages for practical deployment such as the convenience of easy data collection over the telephone. In particular, reliable speaker identification by voice can be extremely useful when other clues to the speaker's identity are either missing or highly ambiguous.

2. 1. Speech and Speaker Recognition

For speaker recognition, the task of the system is either to verify a claimed speaker, or to identify the speaker from some known ensemble. Depending upon the application, speaker recognition can be divided into two related but different subareas as: automatic speaker verification (ASV) and automatic speaker identification (ASI). The aim of speaker recognition is to identify a person from his/her voice by extracting the information from the spoken text, and to answer automatically the question "who is speaking". If it is necessary, the system may also be extended to answer the question "what is said" by extracting different information. In fact, speaker recognition is somewhat related to speech recognition. Ideas in speaker recognition have largely paralleled ideas in speech recognition.

For speech recognition, the task is to understand what is being said rather than who is speaking. In speech recognition, differences due to different speakers in speech signals corresponding to the same text are often perceived as "noise" either to be eliminated by speaker normalisation or more commonly accommodated through the use of different stored spectral patterns for different speakers. But, for speaker recognition, the speech signal must be processed to extract measures of speaker variability instead of analysing it into segments so as to extract the spoken text.

For both kinds of recognition, feature extraction is very important for template matching, and distance measures are common to both applications. However, the reference patterns may be set up with quite different information for speech and speaker recognition. Speaker recognition templates emphasise speaker characteristics while speech recognition templates deal with word information.

2. 2. Identification and Verification

Speaker identification is the process of determining which one of a group of known voices best matches the input voice, whereas speaker verification means determining whether an unknown voice matches the known voice of a speaker whose identity is being claimed. Given a candidate speaker, a speaker identification system attempts to answer the question, "Which speaker (out of a group known to us) is this?", whereas a speaker verification system addresses, "Is he who he says he is?"

In speaker identification an utterance from an unknown speaker has to be attributed, or not, to one of a population of known speakers for whom references are available. More generally, given a total population of N speakers, speaker identification requires choosing which of the N voices known to the system best matches to the pattern of an unknown speaker, as illustrated in Figure 1.

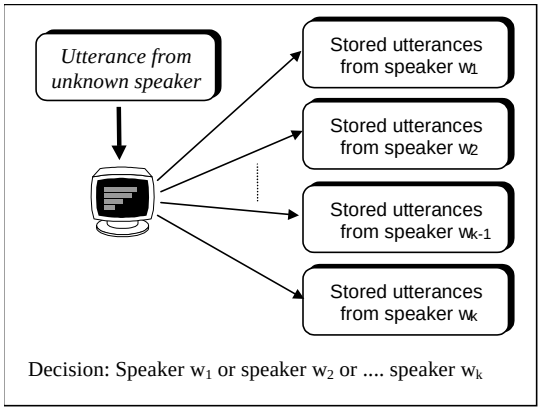

Figure 1. Principle of Speaker Identification

In speaker verification, an identity claim is made by or asserted for the unknown speaker. In this task, the utterance of an unknown speaker together with claimed identity is given and the goal is to determine if the utterance is sufficiently similar to the reference pattern associated with the claimed identity to accept

that claim. In this case, just one comparison of patterns is required regardless of the size of the population, as shown in Figure 2.

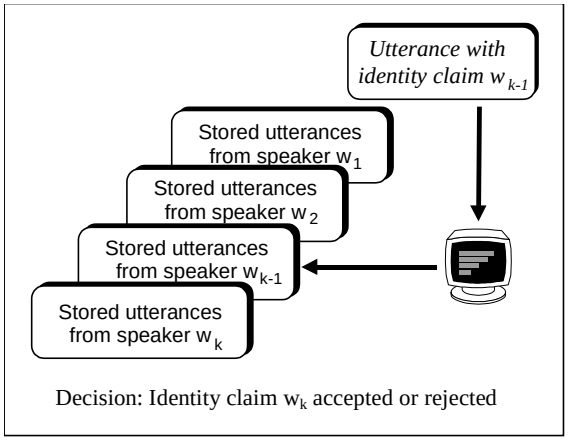

Figure 2. Principle of speaker verification

Although identification and verification tasks have quite a lot in common, the procedures employed in each can be very different. Both verification and identification tasks use a stored database of reference patterns for N known speakers and similar analysis and decision techniques may be employed. However, the most important difference between identification and verification lies in the number of

decision alternatives. It can be said that speaker verification is the simpler task since it requires only a binary decision, namely, that of accepting or rejecting the claimed identity of an utterance and its performance is independent of population size.

In general, speaker recognition can be subdivided into two categories as closed-set and open-set, by dividing the speaker ensemble into two groups of customers (known to the system who claim their true identity) and impostors (unknown to the system for whom the response of the system should be negative), respectively. In a closed-set situation, it is known that the speaker to be identified is one of a population of known (reference) speakers. The speaker that scores best on the test utterance is identified. Naturally, the larger the population, the more difficult is the task. In an open-set test, the speaker to be identified may not be one of this population. If a speaker scores well enough on the basis of a test utterance, then the speaker is accepted as being known. Consequently, in open-set case, an additional decision alternative, "no match", is required. Thus, open-set identification can be thought as a combination of the identification and verification tasks (Doddington, 1985), or speaker verification can be thought as a special case of the open-set identification (Gish and Schmidt, 1994).

In speaker identification, the test speaker may not necessarily be a member of the reference set of speakers. The fact that such membership is unknown is to be ascertained automatically by the system. It means that the decision on identification relies on a comparison of the pattern of the test speaker with each of the individual reference patterns of the speakers from the total population. In verification, the test speaker is assumed to be represented within the reference set. This assumption is held even when an impostor is making a false claim by using the identity of a valid member of the reference set. In the verification task, speakers are assumed to be cooperative and are therefore willing to indicate their claimed identities since they wish to gain access and to have the machine judge whether the claim is correct. In contrast, speakers in the identification task may not wish to be identified and may be uncooperative or may try to disguise their voices, either when making a reference utterance (if they suspect that their voices are being recorded) or when making test recordings. Such disguises or an impostor who is a good mimic may fool the system. In summary, these two problems differ considerably, as Table 1 makes clear:

Table 1. Features That Distinguish Speaker Verification and Identification

VERIFICATION	IDENTIFICATION
a) Speaker is normally cooperative	a) Speaker may be uncooperative
b) Identity Claimed	b) No Claimed Identity
c) Decision:Accept or Reject Claim	c) Decision:Absolute Identification among N
d) One Comparison	d) N Comparisons
e) Mimicry a problem	e) Voice disguise a problem
f) Pr(e)* Independent of N	f) $Pr(e)^* \rightarrow 1$ as $N \rightarrow \infty$
g) System response must be fast	g) System response can be slow
h) Can frequently control channel characteristics	h) Channels may be poor or differing
i) Can usually control signal-to-noise ratio	i) Signal-to-noise ratio may be poor
* Pr(e): Probability of error	

2. 3. Text-Dependent and Text-Independent Recognition

The object of speech communication is to convey a message. This message is the most important information embedded in the speech signal but not the only one. Speech signals also contain information useful for the identification of the speaker, but these two types of information are coded quite differently. However, it is believed that there are no definite acoustic cues specifically and exclusively dealing with speaker identity.

Listeners can normally recognise people from their voices, even though the spoken text is different from

one occasion to another. But in most speaker recognition applications, it is generally required that the texts of the reference and the test utterances be the same, so that corresponding speech events can be compared without the need for coping with the additional variability due to the differences in the texts.

In the literature, speaker recognition is mainly divided into two further subclasses, in terms of the dependency on text as (1) Text-dependent speaker recognition (TDSR) and (2) Text-independent speaker recognition (TISR) (Doddington, 1985; O'Shaughnessy, 1986; Campbell, 1997). TDSR requires the speaker to provide utterances of the same text for both training and testing. A classical approach to TDSR is template matching or pattern recognition, where dynamic time-warping (DTW) methods are usually applied to temporally align the input utterance (testing utterance) and each reference pattern (training utterance) of registered speakers. The reference pattern for each speaker can be represented by using sequences of feature parameters, which depend on the phoneme sequences in the key text. In the TISR case, on the other hand, speakers are not constrained to provide specific texts in training and testing. Since the text spoken by the user can vary each time, it is impossible to represent a reference pattern for each speaker by using sequences of feature parameters, which depend on the phoneme sequences in a key text. Consequently, TISR techniques are primarily based on measurements without reference to a timing index (Markel and Davis, 1979; Soong et all., 1985), and hence, dynamic attributes of spectra may only be exploited in a statistical way (Lin et all., 1994).

Generally speaking, TDSR systems provide better recognition performance than TISR systems especially for short training and testing utterances. Error rates for TISR are considerably higher than for a comparable TDSR case. While TDSR systems typically require 2-3 s of speech for training and for recognition to achieve good results, they usually need much more speech for both training and testing than TISR systems do (Naik, 1990; Peacocke and Graf, 1990).

The use of predefined sentences (or phrase) increases the performance in two ways (Naik and Doddington, 1986):

- Learning problems experienced by the users with the speech material are minimised. The resulting consistency over time enhances system performance.
- The length of the speech material can be conveniently increased yielding better discrimination between true speakers and impostors.

TDSR systems are primarily used in applications (e.g., access control applications) in which the unknown speaker wishes to be recognised and is therefore cooperative. The main reason for TISR is the existence of applications (e.g., forensic and surveillance applications) where there is no guarantee that the speaker will say the same text spoken when the recognition algorithm was trained for him/her. These applications arise when the speaker is uncooperative, perhaps being unaware that recognition is taking place. Theoretically, TISR could be used in any situation where TDSR is applied; the reverse does not hold. TISR is a more general approach to the problem of recognising speakers from their voices.

However, both TDSR and TISR systems have a problem that they can be easily defeated simply by playing back the recorded voice of a registered speaker (Furui, 1994; Matsui and Furui, 1995). To cope with this problem, a small set of words, such as digits, can be used as key words and each user is prompted to utter a given sequence of key words that is randomly chosen every time the system is used (Higgins et all., 1991; Rosenberg et all., 1991). As even this method can be defeated by reproducing key words in a requested order, Matsui and Furui (1993, 1995) have proposed a text-prompted recognition method. In this method, the system prompts each user with a new key text every time the system is used, and accepts the input utterance only when it decides that the registered speaker has uttered the prompted text. Since the vocabulary is unlimited prospective impostors cannot predict the text they will be prompted to say. The system also can reject utterances whose text differs from the prompted text, even if it is uttered by the registered speaker.

3. CLASSIFICATION OF ERRORS AND DECISION RULES

Classification of errors and setting the decision threshold have also crucial importance upon the system performance. It is plausible to classify the errors according to the types of the system used (i.e., identification, verification, open-set, closed-set). In verification task, there are two sources of error whereas error types vary according to closed- set or open-set modes in an identification task. In open-set identification, three kinds of error are possible while in closed-set case, since a match always exist, only one kind of error is possible as shown in Figure 3.

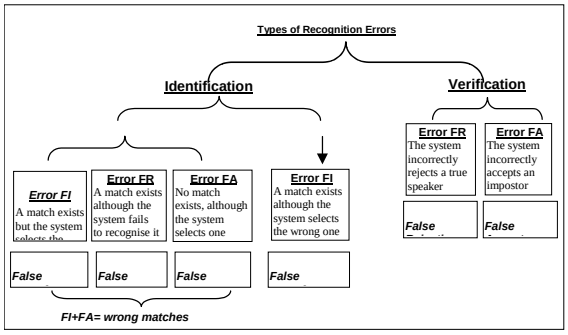

Figure 3. Classification of erros depending on types of speaker recognition system used

In a verification task, a false rejection occurs when the system falsely rejects a true speaker as an impostor, and a false acceptance occurs when an impostor is accepted incorrectly as a customer. In closed-set identification, only a false identification error may occur if system identifies the wrong speaker. But, in open-set, all three types of error may occur. In both open-set identification and verification tasks some form of acceptance/ rejection threshold is required whereas in the closed-set such a threshold is not needed as the "nearest" reference speaker is automatically selected.

The assigned threshold should tolerate trial-to-trial variations, and at the same time ensure a desired level of performance. A "tight" threshold makes it difficult for impostors to be falsely accepted by the system but at the risk of falsely rejecting customers. Conversely, a "loose" threshold enables customers to be accepted consistently, at the risk of falsely accepting impostors. Threshold adjustment creates a trade-off between these two types of errors. Threshold determination should account for the costs of different types of errors the system can commit, e.g., false acceptance error might be more costly than a false rejection error (Campbell, 1997).

A common measure of error is the equal error rate (EER) in which the threshold is set a posteriori so that the two kinds of error rate, the rate of rejecting utterances which should be accepted and the rate of accepting utterances which should be rejected, are equal. The nature of the a posteriori EER measure has been firmly challenged as being unrealistic when related to the problem of selecting a real-life decision threshold, and two methods for determining a priori threshold values have been suggested (Furui, 1981). The first of these involves setting an experimentally determined, fixed-value threshold to remain constant for all claimants. For the second, the optimum threshold is estimated for each claimant, based on his/her reference vector and a set of utterances from other speakers. The results produced using such thresholds show significant deviation from those produced using EER. The EER, however, remains the most common form of performance measurement for automatic speaker verification systems.

Another possible measure, termed the minimal error rate (MER), uses the intersection of the probability density functions produced by the within-speaker and between-speaker distance (Fakotakis et all., 1986). This effectively minimises the sum of the false reject and the false accept rates, thus providing a threshold at which the average error rate will be at a minimum. However, in recent years some researchers, studied the number of the decision alternatives such a discriminant counter (Higgins and Bahler, 1991) and cohort models (Rosenberg et all., 1992).

Most ASV applications require real-time processing, where the system responds immediately to accept or reject a speaker. Such systems may employ a sequential decision procedure, in which borderline decisions are postponed pending further test input. Rather than using a single threshold to accept or reject, two thresholds divide the distance range into three choices: accept if the distance falls below the lower threshold; reject if it's above the higher threshold; and ask for more input if the first distance lies between thresholds (O'Shaughnessy, 1986). Such an approach allows shorter initial test utterances and thus faster response time, while avoiding errors in close cases (Furui, 1981).

4. PERFORMANCE PARAMETERS

Although there are many comparative studies of several speaker recognition systems presented in the literature (Rosenberg and Soong, 1992; Chollet, 1994; Furui, 1994; Matsui and Furui, 1995), there is a major difficulty in comparing studies due to differences in specifications which have vital impact on system performance. Performance is determined in speech tasks by the quality of the speech database evaluated, and reliable performance is often quite easy to achieve if the speech data are carefully controlled. Unfortunately, there are no standard rules to be followed in constructing such a database (Atal, 1976). The differences in database can originate from several sources: number of speakers, type of speaker population, speech material, recording conditions, the time span over which the speech data are collected and the elapsed time between the collection of training and test data. Many researchers stated that the use of benchmark databases for system evaluation has grown in popularity in the last decade in response to the need for meaningful comparative evaluation of systems (Doddington, 1985; Bimbot, et all., 1994;

Campbell, 1997). Otherwise, comparing ASR experiments using different databases is often unreliable. In other words, it is impossible to make serious comparisons of different recognition approaches unless they are evaluated on the same database. Some of the important factors that must be considered in comparing a system with others for an adequate comparison may be:

4. 1. Types of Recognition

Speaker verification or identification. Although these two tasks are quite similar, there is a major difference in the number of decision alternatives according to the type of task. It is difficult to compare between the binary choice verification task and the generally more difficult multiple-choice identification task (Campbell, 1997). The single comparison and the binary choice allow faster computation and less complexity than the compound comparisons and decisions required for speaker identification.

4. 2. Speech Material

Speech input used for speaker recognition could be continuous speech, sentences, single words or phrases, or even (isolated) phonemes. They could be either specifically chosen or arbitrary. Some techniques require more speech input than others to extract speaker-dependent features for recognition. It is also believed that some speech sounds (such as vowels or nasals) carry speaker-specific information better than others (Sambur, 1975). Not only comparison of text-dependent and text-independent systems but also comparison of text-dependent systems is difficult while different systems use different protocols such as, type of voice password, decision strategy, training and update methods, etc. Text-independent systems are less constrained by some of these issues but type of speech material and amount of testing and training data also vary widely among the systems under development (Naik, 1994).

4. 3. Speaker Ensemble

The composition and characteristics of the speaker population are important parameters that should be considered carefully. Selected speaker ensembles may include many different kinds of people. Speakers could be cooperative or uncooperative, trained or untrained, child or adult, native speakers of the language or foreigners, male, female or mixed-set, etc. The system can be evaluated with either only the customers or both the customers and impostors. Many studies have used only male speakers because of the difficulties associated with analysis of female speech, which are well known (Junqua and Haton, 1996). Differences in speakers' accents and speaking styles are also very important.

4. 4. Population

One factor which defines the difficulty of the speaker identification task is the size of the speaker population. Ideally, the test population for speaker recognition studies should be as large as can be managed. But in practice, limitations such as storage capacity and speed of access to reference patterns as well as the need for collecting data from a large number of speakers can give many practical problems. In fact, population size is a critical performance parameter for speaker identification while the performance for speaker verification is unaffected by population size (Doddington, 1985; Naik, 1990). In the case of identification the reliability of recognition decreases as the number of speakers increases, whereas the recognition rate for verification is independent of the number of speakers. Hence, verification systems may serve practically any number of users. The distinction between identification and verification has practical consequences. The similarity of the speakers in the population also must be considered, since a set of speakers with dissimilar voice characteristics usually yields higher recognition performance than a more similar set of speakers (Reynolds and Rose, 1995).

4. 5. Types of Error

There are strong differences of meaning between false acceptance and rejection rates, depending on the way an impostor is defined: whether an impostor claims an identity at random or the identity of its closest neighbour in the test corpus makes a lot of difference in the performance evaluation. For speaker verification, some authors test all other speakers in the databases as impostors, while some others test only the second closest. In particular, some results take into account the a priori probability of impostors (generally small), while some others do not, and some suppose that the impostor knows whose voice is closest to his, while others do not make this assumption. A standard protocol for impostor testing is also necessary.

4. 6. Training/Testing

Speaker recognition systems may be evaluated under two conditions: matched or unmatched conditions between training and testing. A matched condition corresponds to training and testing under identical system conditions, i.e., frequency response, microphone, SNR, reverberation, etc. An unmatched condition corresponds to training and testing under

different system conditions. Generally, the matched condition gives higher identification scores than the unmatched condition. 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. The effects of a time difference between reference and test data collection sessions is also important. In speaker verification, the performance of a system asymptotically approaches a stable level after about 10 to 15 sessions per speaker, assuming that some form of adaptation of the speaker model is used. Hence, speech should be recorded in several sessions, over a duration of 3 to 6 months, at different times of day (Bimbot, et all., 1994; Naik, 1994).

4. 7. Environment

Environmental conditions are of crucial importance to the performance of a system. Recording environment and equipment are of particular concern. Was the recording place quiet? Was it an ordinary room or a special anechoic chamber? What kind of microphone and recording machine have been used? Were speech data recorded over the telephone line or not? To make a reasonable comparison between different speaker recognition systems, such environmental conditions should be the same or, at least, fairly close together.

4. 8. Implementation

The type and capacity of the computer used for evaluation of the system is also important. Dealing with a large population requires large storage capacity. Speed of access to reference patterns also depends on computer capacity.

5. APPLICATIONS OF SPEAKER RECOGNİTİON

Automatic speaker recognition systems can be applied in three areas:

1) forensic investigations (associating a person with a voice in police work),
2) security systems for confirmation of identity (verifying a person's identity prior to admission to a secure facility or a transaction over the telephone), and
3) military applications.

In the forensic field such as criminal investigations voice patterns of an unknown speaker are compared with voice patterns of suspects to decide whether or not a match can be obtained. The person to be identified will try not to be recognised if he has perpetrated a criminal action. For this reason the voice will often be disguised, for example for telephone calls (blackmail attempts, bomb threats, etc.). Also, it can not be expected that the speaker will be cooperative in providing his voice after he is apprehended. On the other hand, in this kind of application, it is usually not possible to exercise control over the text of the test utterance or the recording and transmission conditions. Some of the speaker recognition techniques may not be suitable for applications such as these which are associated with large amounts of uncontrolled variability. Nevertheless, with Kersta (1962) voiceprint methods, or more generally spectrograms, came to be used for speaker identification, especially in forensic applications. The ability to identify speakers via voiceprints has been of particular interest in forensic work. Despite some evidence to the contrary, most researchers feel that spectrogram reading has not been demonstrated to identify speakers reliably. Experts seem to achieve a certain degree of ability to match reference spectrograms to test ones by the same speaker, but performance often degrades substantially if speakers disguise their voices.

Speech spectrograms, when used for voice identification, do not correspond with fingerprints, because of basic differences in the sources of the patterns. As an example, fingerprint patterns are a direct representation of anatomical characteristics. In contrast, vocal anatomy is not represented in any direct way by voice spectrograms.

Bolt et all., (1970) have reported a newer method of voice identification which uses visual comparison of the graphic patterns resulting from a gross acoustic analysis using the sound spectrograph. Not all details of the acoustic patterns are presented in this graphic display; moreover, the display is designed to emphasise those features that characterise the words of the spoken message. Speech-sound spectrograms of this type are the primary material used forensically for voice identification. They published a lengthy critique of the spectrographic method, in which the following conclusions were stated:

1. It is possible, to a limited extent, to identify voices from spectrograms,
2. Spectrograms do not facilitate distinguishing between speaker-dependent and message-dependent features.,
3. Similarities and differences among spectrogram patterns may arise from many different sources and can be misleading,
4. Spectrograms are not the same as fingerprints, since speech patterns change over time, while fingerprints do not,
5. Research has yielded highly inconsistent results, depending on details of the methods used,

6. Success in a criminal trial depends on many factors and does not imply validity of spectrographic identification techniques,
7. Rigorous experiments simulating the conditions found in law-enforcement applications have not been made.

A more suitable application for automatic speaker recognition techniques is a security device to control access to buildings or information (e.g., Texas Instruments Access Control System). There are not many systems in which a significantly large sample of speakers and utterances have been used for evaluation, or in which the collection and recording of the sample has been accomplished under "real-world" conditions. Most studies, in fact, can be considered preliminary laboratory investigations of particular speaker recognition techniques with no special claims to be "real-world" systems operating outside the laboratory. When facilities or information must be secured from access by unauthorised persons, speaker recognition offers inevitable advantages. In this kind of application it can be expected that the speakers are cooperative. They respond to instructions and try to control variability. Moreover, in such an application, the user who desires the machine to perform a useful service for him is willing to indicate his claimed identity and to have the machine decide if the claim is correct.

In addition to its use as a security device, speaker recognition could be largely used in electronic banking such as commercial banking systems for telephone-based transactions, credit card verifications, and voice mails (e.g., The AT&T Bell Labs Automatic Speaker Verification System, Siemens Verification System, VERIFIER by ENSIGMA Ltd., and BT home banking trial with Royal Bank of Scotland). Applications can be extended to: voice-directed installation of telephone equipment, verification by voice of a credit customer or of an individual requesting readout of privileged information, and voice-controlled services such as automatic booking of travel reservations, remote access to computers via modems on dial-up telephone lines, and one can even configure an answering machine to deliver personalised messages to a small set of frequent callers. Access to cars, buildings, important information, bank accounts and other services may be voice controlled in the future as a result of advances in digital signal processors and speech technology that have made possible the design of fast, cost effective, high-performance speaker recognition systems.

Another area of speaker recognition application is primarily of interest to the military. Automatic speaker recognition, perhaps in conjunction with keyword recognition, can allow more complete coverage of enemy communications, allowing identification of dangerous situations before they occur.

6. REFERENCES

Atal, B. S. 1976. Automatic Recognition of Speakers From Their Voices. Proc. IEEE. 64 (4). 460-475.

Bennani, Y. and Gallinari, P. 1994. Connectionist Approaches for Automatic Speaker Recognition. Proc. ESCA Workshop. 95-102.

Bimbot, F., Chollet, G. and Paoloni, A. 1994. Assessment Methodology for Speaker Identification and verification systems: an overview of SAM-A Esprit project 6819-Task 2500. Proc. ESCA Workshop. 75-82.

Bolt, R. H., Cooper, F. S., David, E. E., Denes, P. B., Pickett, J. M., Stevens, K. N. 1970. Speaker identification by speech spectrograms: A scientists' view of its reliability for legal purposes. JASA, 47(2), 597-612.

Campbell, J. P. 1997. Speaker Recognition: A Tutorial. Proc. IEEE, Vol. 85, No. 9, 1437-1463.

Chollet, G. 1994. Automatic Speech and Speaker Recognition: Overview, Current Issues and Perspectives. In Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State-of – Art and Future Challenges. Ed. E. Keller. John Wiley and Sons.

Doddington, G. R. 1985. Speaker Recognition-Identifying People by Their Voices. Proc. IEEE, 73 (11), 1651-1664.

Fakotakis, N., Dermatas, E. and Kokkinakis, G. 1986. Optimal Decision Threshold for Speaker Verification. 3rd European Signal Processing Conf. EUSIPCO'86. 585-588.

Furui, S. 1981. Cepstral Analysis Technique for Automatic Speaker Verification. IEEE ASSP-29, No. 2, 254-272.

Furui, S. 1994. An Overview of Speaker Recognition Technology. Proc. ESCA Workshop. 1-9, Martigny, April 5-7, 1994.

Gish, H. and Schmidt, M. 1994. Text-Independent Speaker Identification. IEEE Signal Proc. Mag. Vol.11, No.4, 18-32, October 1994.

Higgins, A. L., Bahler, L. 1991. Text-independent Speaker Verification by Discriminator Counting. IEEE Pro. ICASSP'91. Vol.1. 405-408.

Higgins, A. L., Bahler, L. and Porter, J. 1991. Speaker Verification Using Randomized Phrase Prompting. Digital Signal Processing. Vol. 1. 89-106.

Jesorsky, P. 1978. Principles of Automatic Speaker Recognition. in Speech Communication With Computers, Ed. Bolc, L., Macmillan.

Junqua, J-C. and Haton, J. P. 1996. Robustness in Automatic Speech Recognition: Fundamentals and Applications, Kluwer Academic Publishers, 1996.

Kersta, L.G. 1962. Voiceprint Identification. Nature, 196, 1253-1257.

Lastrucci, L., Gori, M. and Soda, G. 1994. Neural Autoassociators for Phoneme-based Speaker Verification. Proc. ESCA Workshop. 189-192, Martigny, April 5-7, 1994.

Lin, Q., Jan, E. E. and Flanagan, J. 1994. Microphone Arrays and Speaker Identification. IEEE SAP, Vol. SAP-2, No. 4, 622-629.

Markel, J. D. and Davis, S. B. 1979. Text-Independent Speaker Recognition From a Large Linguistically Unconstrained Time-spaced Data Base. IEEE ASSP-27 (1). 74-82.

Matsui, T. and Furui, S. 1993. Concatenated Phoneme Models for Text-variable Speaker Recognition. IEEE Proc. ICASSP'93. 391-394.

Matsui, T. and Furui, S. 1995. Speaker Recognition Technology. NTT Review, Vol. 7, No. 2, 40-48, 1995.

Naik, J. M. and Doddington, G. R. 1986. High Performance Speaker Verification Using Principal Spectral Components. IEEE Proc. ICASSP'86. 881-884.

Naik, J. M. 1990. Speaker verification: A tutorial. IEEE Comm. Mag. Vol. 28, No. 1, 42-48, January 1990.

Naik, J. M. 1994. Speaker Verification over the Telephone Network: Databases, Algorithms and Performance Assessment. Proc. ESCA Workshop. 31-38.

O'Shaughnessy, D. 1986. Speaker Recognition. IEEE ASSP Magazine, 4-17, October 1986.
O'Shaughnessy, D. 1990. Speaker Recognition. in Speech Communication: Human and Machine, Chapter 11, Addison-Wesley Publishing, 1990.

Peacocke, R. D. and Graf, D. H. 1990. An Introduction to Speech and Speaker Recognition. IEEE Trans Computer. Vol. 23. No. 8. 26-33.

Reynolds, D. A. and Rose, R. C. 1995. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE SAP-3, No. 1, 72-83, January 1995.

Rosenberg, A. E., Lee, C. H. and Gökçen, S. 1991. Connected Word Talker Verification Using Whole Word HMMs. IEEE Proc. ICASSP'91. 381-384.

Rosenberg, A. E. and Soong, F. K. 1992. Recent Research in Automatic Speaker Recognition. in Advances in Speech Signal Processing. Eds. Furui, S. and Sondhi, M. Marcel Dekker, 1992.

Rosenberg, A. E., DeLong, J., Lee, C.-H., Juang, B. H. and Soong, F. K. 1992. The use of Cohort Normalized Scores for Speaker Verification. Proc. Int. Conf. Spoken Language Processing. 599-602.

Sambur, M. R. 1975. Selection of Acoustic Features for Speaker Identification. IEEE ASSP-23, 169-176.

Soong, F. K., Rosenberg, A. E., Rabiner, L. R. and Juang, B. H. 1985. A Vector Quantization Approach to Speaker Recognition. Proc. IEEE ICASSP'85, Tamoa, FL 387-390, 1985.

Woodward, J. D. 1997. Biometrics: Privacy's foe or Privacy's Friend? Proc. IEEE, Vol. 85, No. 9, 1480-1492, September 1997.