Erciyes Üniversitesi Fen Bilimleri Enstitüsü Dergisi, **18** (1-2), 34-43, 2002

PRACTICAL ASPECTS OF BUILDING A SPEECH DATABASE FOR SPEAKER RECOGNITION SYSTEMS

Figen ERTAŞ

Uludağ Üniversitesi Mühendislik Mimarlık Fakültesi, Elektronik Mühendisliği Bölümü Görükle/BURSA

Abstract: Speech databases are of vital importance in many area of speech research. In particular, the evolution of current automatic speech and speaker recognition systems has been strictly related to the availability of large corpora of speech. The success of developing practical systems heavily depends on the use of proper databases. However, there are difficulties in the definition of universal standards for the development of speech databases. A certain difficulty is the presence of many variabilities in the speech signals. In this paper, sources of variability in speech and speakers that affect the system performance are briefly discussed with regard to building a speech database. A comparative discussion on most popular databases is also given.

Key words: Speech, speaker, environment, variability, speech database

KONUŞMACI TANIMA SİSTEMLERİ İÇİN SÖZ VERİTABANI OLUŞTURMANIN PRATİK YÖNLERİ

Özet: Söz veri tabanları, söz araştırmalarının bir çok alanında hayati öneme sahiptir. Özellikle otomatik söz ve konuşmacı tanıyan sistemlerin gelişimi, geniş kapsamlı söz veri tabanlarının elde bulunmasına bağlı olmuştur. Pratik sistemlerin geliştirilmesindeki başarı, büyük oranda uygun veri tabanı kullanılmasına bağlı olmaktadır. Bununla beraber, söz veri tabanı oluşturmak için ortak bir standart tanımlamada büyük zorluklar bulunmaktadır. Bu zorluklardan biri, konuşma işaretlerindeki mevcut değişimlerdir. Bu makalede, konuşma işaretindeki değişimler etraflıca tartışılmış ve en bilinen veri tabanları için karşılaştırmalı bir inceleme yapılmıştır.

Anahtar kelimeler: Konuşma, konuşmacı, çevre, değişim, söz veri tabanı

1. Introduction

A speech database is a collection of recorded speech accessible on a computer and supported with the necessary annotations and transcriptions. There are three categories of speech databases currently available: the *analytic-diagnostic*, *generic*, and the *specific*. The first type is used to improve our knowledge of the fundamental linguistic and phonetic elements of speech. The second type includes non-specific vocabularies, while the third collects target-specific speech.

Nevertheless, recently, there has been more interest in the development of databases containing "*natural speech*" recorded in situations of everyday conversations about topics chosen by the speaker. In contrast, "*laboratory speech*" is recorded in controlled situations and produced in a style that is much more formal. Laboratory speech can include different pronunciation styles. "*Read speech*" is produced by speakers with different training and an awareness of the topic to be read. Instead, "*spontaneous speech*" is collected from monologues, dialogues, or from simulation of human-computer interaction, with a human operator simulating the computer response. This last

technique is referred to as the "*Wizard of Oz*".

In this paper, we discuss the sources and types of variations in speech signals, and introduce parameters that must be taken into account in preparing speech databases for speaker recognition applications. We also present a comparative discussion on the most commonly used speech databases for both speech and speaker recognition.

2. Sources of Speech Varriety

Speech signals convey information at several levels. Primarily, they convey the words (or the message) that was said but, at a secondary level, they convey information about the identity of the speaker. In addition, speech signals include clues to the physical and emotional state of the speaker (manner and mood of the speaker, e.g., anger, fear, etc.), state of the speaker's health, class of the speaker (man, woman or child, his/her sociological background, his/her geographical origin), and the recording environment, as illustrated in Figure 1. In other words, many different pieces of information are carried simultaneously in a single acoustic stream as a speech signal. Thus, there are large variabilities in the speech signal between speakers and, more importantly, significant variations from instant to instant for the same speaker and text.

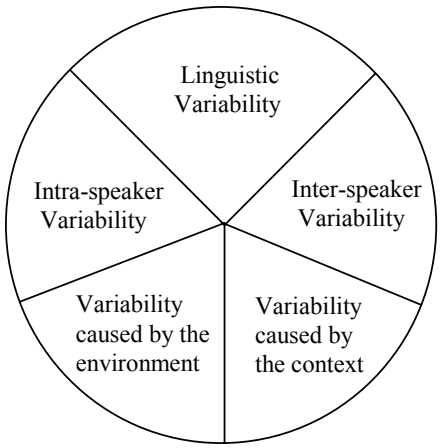

Figure 1 Classification of speech variability in five categories *(Junqua and Haton, 1996)*

3. Sources of Variability Between Speakers

When two persons speak the same utterance, their articulation is similar but not identical; thus, spectrograms of these utterances will be similar but not just the same. Even when the same speaker utters the same word on two different occasions, there are also similarities and differences. There are many reasons why some aspects of the sound pattern of an utterance are different on different occasions. For different speakers, the voices of two persons differ due to the physical differences between their vocal anatomies and the manner in which they use them during speech production. The possibility of two people having all vocal organs the same size and coupled identically seems remote. A greater factor in determining voice uniqueness is the way in which the articulators are used during speech and, again, the chance that two people would have the same dynamic use-patterns for their articulators would be remote (Kersta, 1962). Also, a single speaker may show considerable variation in their use from one utterance to another.

We do not have yet an insight of which speech features are likely to be invariant for a given speaker, and which are

likely to show variation from one speaker to another. The task is to tease out from the speech patterns those features which correspond to the speaker's vocal anatomy and his/her habits of forming speech sounds, since these might characterise him/her as a speaker. Unfortunately, it is not easy to define a set of simple and physically meaningful parameters for speech conveying information solely about the speaker. Current results indicate that speaker-dependent and text (utterance)-dependent information in speech signals are combined in a complex mode at the acoustic level and nearly all acoustic parameters are speaker-dependent to some extent.

Broadly speaking, there are two main sources of variation among speakers: anatomical (or physiological) differences and learned (or behavioural) differences which lead to two types of useful features as *inherent* and *learned* features. The anatomical differences from speaker to speaker relate to the lengths (sizes) and shapes of the components of their vocal tracts: larynx, pharynx, tongue, teeth, and the oral and nasal cavities. For example, a shorter vocal-tract length results in higher resonant (formant) frequencies. Variations in the size of vocal tract cavities produce differences in the characteristic resonances of the spectrum of speech signals. Variations in the size of vocal cords are associated with changes in the average pitch or fundamental frequency of voiced speech. Variations in the velum and size of the nasal cavities produce spectral differences in nasalized speech sounds. As a result of the natural physiological variations, inherent features are relatively fixed for a speaker and can be affected by health conditions (e.g., colds that congest the nasal passages).

Learned features, i.e., the way a speaker talks, are not given by nature but are gained through learning to use his/her speech mechanism and practical use of a language. Learned features might be useful for distinguishing people with similar vocal mechanisms. Such differences reveal themselves in the temporal variations of speech peculiarities of different people and cause differences in the dynamics of the vocal tract such as the rate of formant transitions and coarticulation effects. They also affect speaking rate, stress, and melody. As inherent features are less sensitive to counterfeit than learned features, impostors generally find it easier to fool recognisers that are based on learned features than those using inherent features (O'Shaughnessy, 1986).

Since sources of variability are numerous, two of the main sources of speaker variation are directly related to the physiological and psychological state of the speaker and to the speech communication goal. In the subsequent sections, inter-speaker and intra-speaker variabilities are briefly discussed.

3.1 Intra-Speaker Variability

The most important factor affecting automatic speaker recognition performance is variations in speech characteristics that occur from session to session for the same person. The time span over which speech is recorded is of crucial importance to the system performance (O'Shaughnessy, 1986). It has been shown spectrographically (Kersta, 1962; Bolt *et all*, 1970), that speakers usually sound the same utterance very differently from each other, which is good for distinguishing speakers, but the real problem is that a single speaker also often sounds very different from time to time (Doddington, 1985).

Variations may also arise from differences in recording, transmission conditions, and voice. But the most significant is the variation produced by the same speaker, which can be voluntary or involuntary. These variations may become so large as to render any speaker recognition decision completely unreliable. Even under the same conditions, speakers cannot repeat an utterance precisely the same way from trial to trial. This phenomenon is called "*intra-speaker variability*". Intra-speaker variations occur within different speech utterances of a single speaker or even though the texts of two utterances are the same. The differences in speaking rates, the emotional state and health of

the speaker and speaking style, which often changes considerably over time, could be the possible factors for such variations.

It is likely that speakers can change their voice quality, their speaking rate, their fundamental frequency or even their articulation patterns. When small changes occur in their articulation pattern, there can be big differences at the acoustic level. Speakers can also change their pronunciation, affecting the spectral characteristics of individual speech sounds by varying the amount of vocal effort substantially from utterance to utterance (Rosenberg and Soong, 1992). The long-term spectrum is generally assumed as one of the most reliable cues to voice quality. The natural variation of the long-term spectrum, which is associated with both changing speaking behaviour and ageing, has been found to be speaker-dependent (Harmegnies and Landercy, 1988). An automatic speech recognition system should therefore adapt to natural and expected modifications in speech signal characteristics due to all these type of variations.

Furthermore, external acoustic conditions, such as multiple speakers or environment changes, can also affect the quality of the speech signals being analysed and cause intra-speaker variability. For example, background noise or stress conditions produce an increase in the speakers' vocal effort and hence a variation of speech production, which causes acoustic-phonetic variations.

Since reference and test data recorded in the same session are more likely to be highly correlated than those recorded in different sessions, they would mislead as to system performance if used for evaluation. It has been shown that performance usually decreases when the period between training and testing sessions increases (Furui, 1981). Practical applications generally use test data that were collected much later than reference data. However, it would improve the system performance if the reference data were periodically updated.

3.2 Inter-Speaker Variability

When the same utterance is spoken by different speakers, differences are perfectly observable to the listeners, and are known as "*inter-speaker variability*". They mainly originate from physiological differences such as length and shape of the vocal tract, physiology of the vocal folds, shape of the nasal tract, which generate acoustic variability, and from learned differences in the use of the speech mechanism. For example, it is well known that the length and the shape of the vocal tract varies between speakers, and therefore, the consequent formant frequencies also vary for different speakers, e.g., a shorter vocal tract length usually renders higher formant frequencies. Also, there are some basic differences between male and female speakers such as formant frequencies, which are about 15% higher for female speakers than those for male speakers (Vaissiere, 1985). In addition, a person's vocal mechanism grows and changes with age, and hence an utterance generated by an adult would be expected to be acoustically different from the same utterance generated by a child. Junqua (1993) reported that when speech is produced in noise, the compensation methods adopted by each speaker are different. As yet, neither intra-speaker variability nor inter-speaker variability has not been quantified or correlated with specific acoustical parameters of the speech signal (Tosi *et all*, 1972).

4. Environment variability

In system designs, careful attention should be paid to the impacts of the environment on performance. The environment in which speech is produced and recorded plays a crucial role and affects its production, perception and acoustic representation. Junqua and Haton (Junqua and Haton, 1996) classified the environmental factors causing

this variability in two categories:

- *static* (room acoustics and reverberation, recording tools and the speaker personality and physical characteristics), and
- *dynamic* (background noise, emotion, stress, and microphone placement).

Most experiments have been conducted in noise-free and ideal environments. The sites for real applications often present adverse conditions which can drastically alter the system performances. Dealing with reasonable noise and distortions of the speech signal due to environmental conditions, maintaining a reasonable level of performance irrespective of the microphone used, and achieving performance robustness against variability in the transmission line characteristics (e.g., telephone) constitute a wide area of research topics in speaker recognition. Although, some significant results have indeed been obtained under laboratory or near laboratory conditions, these performances fall rapidly once systems are confronted with realistic conditions or have to cope with untrained users.

5. Design of Speech Database

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. The differences in database can originate from several sources:

- type of speech material,
- type of speaker population,
- number of speakers,
- 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; Naik, 1994; Campbell, 1997). Otherwise, comparing speech 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:

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 (text-dependent or text-independent). 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).

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.

Population: One factor which defines the difficulty of the speaker identification task is the size of the speaker population. 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).

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.

Training/Testing: 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 (Naik, 1994). 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).

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.

6. Comparision of Well-Known Speech Databases

In the last decade, a number of standardised speech databases has significantly contributed to the evaluation of speech technology. It is imperative that standardised databases be developed and shared in the speech research community to measure progress reliably and evolve new techniques to improve current methods. Several databases have been introduced for development and evaluation of different speech and speaker recognition systems (Naik, 1994; Campbell, 1997; Gish and Schmidt, 1994; Godfrey *et all*, 1994; Reynolds, 1994; 1995). Each one has its own design characteristics which make it more appropriate for certain types of research. An overview of the most prominent databases for both speech and speaker recognition systems is given in Table 1. The table contains the organisation where the database was collected, speech material used, number of speakers, acoustic environment where speech data was recorded (over the telephone line, ordinary room, or an anechoic chamber), sampling rate/format, and the purpose for which database was collected.

Table 1. Characteristics of Main Databases

NAME	organization	speech material	Number of speakers	recording environment	data source	sampling rate (kHz) & sampling format	purpose
TIMIT	recorded at TI, transcribed at MIT, prepared for CD-ROM by NIST, sponsored by DARPA	6300 sentences (630 speakers each reading 10 sentences)	**630** 438 Male + 192 Female	Quiet Room	microphone	16 1 channel 16-bit linear	speech recognition
NTIMIT	developed by NYNEX	6300 sentences (630 speakers each reading 10 sentences)	**630** 438 Male + 192 Female	telephone (TIMIT transmitted through a telephone handset)	telephone	16 1 channel 16-bit linear	speech recognition
CTIMIT	collected by VCI at SPCOT, sponsored by AE&T Division	6300 sentences (630 speakers each reading 10 sentences)	**630** 438 Male + 192 Female	cellular telephone	telephone	8 1 channel 16-bit linear	speech recognition
HTIMIT	re-recording of a subset of the TIMIT through different telephone handsets	3840 sentences (10 TIMIT sentences for 384 speakers)	**384** 192 Male + 192 Female	telephone (background noise and channel variations included)	telephone	8 1 channel 16-bit linear	speaker ID **&** recognition
KING-92	collected at ITT in 1987, reprocessed in 1992 by the LDC at the University of Pennsylvania	10 sessions/ speaker each consists of 30s actual speech of the person speaking on assigned topics	**51** All Male	Quiet Room	microphone (mounted on the telephone handset)	8 (original at 10) 1 channel 16-bit linear	speaker ID
YOHO	collected by ITT under a US Government contract	1932 phrases (138 speakers each reading a sequence of three two-digit numbers)	**186** 156 Male + 30 Female	Real-world office	microphone	8 1 channel 16-bit linear compressed	speaker ver.
SWITCH BOARD (Release 2)	collected at TI, sponsored by DARPA	2400 two-sided telephone conversations among 543 speakers	**543** Male + 241 Female	telephone (long distance telephone line)	telephone	8 2 channel ulaw	speaker ID **&** speech recognition
SPIDRE	derivative subcorpus of SPIDRE, selected for speaker ID	360 dialogues (45 speakers each done two-sided 4 conversations)	**45** Male + 22 Female	telephone (noise, cross-talk and handset variations included)	telephone	8 2 channel ulaw	speaker ID

The **TIMIT** corpus of read speech has been designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. TIMIT contains broadband recordings of

630 speakers (438 male and 192 female) of 8 major dialects of American English, each reading 10 phonetically rich sentences. Speech data were recorded from a close-talking microphone, bandlimited to 8 kHz. There is no intersession variability, no acoustic noise (recorded in a sound booth), no microphone variability, and no distortion. The TIMIT corpus includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16kHz speech waveform file for each utterance. Corpus design was a joint effort among the Massachusetts Institute of Technology (MIT), SRI International (SRI) and Texas Instruments, Inc. (TI). The speech was recorded at TI, transcribed at MIT and verified and prepared for CD-ROM production by the National Institute of Standards and Technology (NIST).

The **NTIMIT** corpus was developed by the NYNEX Science and Technology Speech Communication Group to provide a telephone bandwidth adjunct to TIMIT. NTIMIT was collected by transmitting all 6300 original TIMIT recordings through a telephone handset and over various channels in the NYNEX telephone network and redigitizing them. In order to calibrate the transmission characteristics of the various channels, stationary 1 kHz and frequency-sweeping tones were also recorded for each of the transmission channels.

The **CTIMIT** corpus is a cellular-bandwidth adjunct to the TIMIT. The corpus was contributed by Lockheed-Martin Sanders to the LDC for distribution on CD-ROM media. The CTIMIT read speech corpus has been designed to provide a large phonetically labeled database for use in the design and evaluation of speech processing systems operating in diverse, often hostile, cellular telephone environments. CTIMIT was collected by members of the Voice Communication Initiative (VCI) at Lockheed-Martin Sanders' Signal Processing Center of Technology (SPCOT) as part of internal R&D efforts, with additional sponsorship from the Wireless Communications Group in the company's Advanced Engineering and Technology (AE&T) Division.

The **HTIMIT** corpus is a re-recording of a subset of the TIMIT corpus through different telephone handsets. The aim was to create a corpus for the study of telephone transducer effects on speech which minimized confounding factors, such as variable telephone channels and background noise. HTIMIT was created by playing 10 TIMIT sentences from 192 male and 192 females through a stereo loudspeaker into different transducers positioned directly in front of the loudspeaker and digitizing the output from the transducers. Ten transducers (telephone handsets) were used. In order to obtain some diversity with a limited number of handsets, handsets were selected to have variable sound characteristics, transducer designs. HTIMIT offers the ability of studying handset transducer effects on speech recognition systems.

The **KING** corpus is designed principally for closed set experiments in text-independent speaker identification or verification over toll-quality telephone lines, although the single-sided collection format does not permit simulation of real telephone traffic It was collected partly in New Jersey and partly in San Diego in 1987, the version now available at the LDC is a 1992 reprocessing of these original recordings. It contains recorded speech from 51 male speakers in two versions, which differ in channel characteristics: one from a telephone handset and one from a high-quality microphone. There are 26 San Diego speakers and 25 New Jersey speakers. For each speaker and channel there are ten files, corresponding to sessions of about 30 to 60 s duration each. The interval between sessions varies from a week to a month.

The **YOHO** database supports only text-dependent speaker verification research such as used in “secure access” technology. The large number of speakers (156 male, 30 female, total of 186) and the systematic set of impostor utterances makes it ideal for this type of research (Godfrey *et all*, 1994). The data was collected in 1989 by ITT under a US Government contract. The syntax used in the YOHO database is “combination lock” phrases. Each

phrase was a sequence of three two-digit numbers (e.g., twenty-six, seventy-one, forty-five). There is no handset or channel variation. YOHO database collected at 8 kHz sampling with 3.8 kHz analog bandwidth over 3 month period in a real-world office environment (Che and Lin, 1995).

The **SWITCHBOARD** corpus, originally collected by Texas Instruments with funding from the Advanced Research Projects Agency (ARPA), is designed to support several types of speech and language research. Its variety of speakers, speech data, telephone handsets, and recording conditions make it a rich source for speaker verification experiments of several kinds (Godfrey *et all*, 1994). It consists of spontaneous conversational speech recorded over long-distance telephone lines from 543 speakers (302 male, 241 female) representing all regions of the U.S. But, the amount of data (about 240 hours of speech, about 3 million words, over 12 GB of data) imposes certain limitations for evaluation purposes, as it is unlikely that no two investigators can do similar experiments independently. For this reason, its derivative subset, the **SPIDRE** (**Sp**eaker **Id**entification **Re**search) corpus, with manageable size and special attention to telephone handset variation has been created.

SPIDRE corpus contains training and testing data for experiments in closed or open set recognition or verification. Combining the two sides of the conversations also permits speaker change detection, or speaker monitoring, experiments. There are 45 "target" speakers (23 male and 22 female); 4 conversations from each target are included, of which 2 are from the same handset. There are also 100 calls in which no target appears. Since all conversations are two-sided, this results in 180 target sides and 380 nontarget sides.

7. Conclusion

In recent years, there have been increasing interest in deploying speaker verification systems, particularly in the cellular telephone network, where security has become a prime issue. Databases are necessary tools for the development and research in speech processing, and need to serve the purpose of objectively evaluating the merits of each system. In this paper, we have discussed the sources and types of variations in speech signals, and introduced parameters that must be taken into consideration in building speech databases for speaker recognition applications. The impact of parameters are discussed in terms of system performance. We have also presented a comparative discussion on the most commonly used speech databases for both speech and speaker recognition applications.

References

1. Bimbot, F., Chollet, G. and Paoloni, A., Assessment methodology for speaker identification and verification systems: an overview of SAM-A Esprit project 6819 - Task 2500. Proc. ESCA, April 5-7, 1994, Martigny, 75-82, 1994.
2. Bolt, R. H., Cooper, F. S., David, E. E., Denes, P. B., Pickett, J. M. and Stevens, K. N., Speaker identification by speech spectrograms: A scientists' view of its reliability for legal purposes. JASA, 47(2), 597-612, 1970.
3. Campbell, J. P., Speaker recognition: A tutorial. Proc. IEEE, Vol. 85, No.9, September 1997, 1437-1463, 1997.
4. Che, C. W. and Lin, Q., Speaker recognition using HMM with experiments on the YOHO database. ESCA, Eurospeech'95, 4th European Conf. on Speech Comm. & Tech.,Vol.1, 625-628, 1995.
5. Doddington, G. R., Speaker recognition-identifying people by their voices. Proc. IEEE, 73(11), 1651-1664, 1985.
6. Furui, S., Comparison of speaker recognition methods using statistical features and dynamic features. IEEE Trans. Acoust., Speech and Signal Proc. ASSP-29, No.3, June 1981, 342-350, 1981.

7. Gish, H. and Schmidt, M., Text-independent speaker identification. IEEE Signal Processing Magazine, Vol.11, No.4, October 1994, 18-32, 1994.

8. Godfrey, J., Graff, D. and Martin, A., Public databases for speaker recognition and verification. Proc. ESCA, April 5-7, 1994, Martigny, 39-42, 1994.

9. Harmegnies, B. and Landercy, A., Intraspeaker variability of the long-term speech spectrum. Speech Communications, 7(1), 81-86, 1988.

10. Junqua, J. C., The Lombard reflex and its role on human listeners and automatic speech recognizers. JASA, 93(1), 510-524, 1993.

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

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

13. Naik, J. M., Speaker verification over the telephone network: databases, algorithms and performance assessment. Proc. ESCA, April 5-7, 1994, Martigny, 31-38, 1994.

14. O'Shaughnessy, D., Speaker recognition. IEEE ASSP Magazine, October 1986, 4-17, 1994.

15. Reynolds, D. A., Experimental evaluations of features for robust speaker identification. IEEE Trans. on SAP, Vol. SAP-2, No.4, October 1994, 639-643, 1994.

16. Reynolds, D. A., Large population speaker identification using clean and telephone speech. IEEE Signal Proc. Letters, Vol.2, No.3, March 1995, 46-48, 1995.

17. Reynolds, D. A. and Rose, R**.** C., Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Trans. Speech and Audio Proc. SAP-3, No.1, January 1995, 72-83, 1995.

18. Rosenberg, A. E. and Soong, F. K., Recent research in automatic speaker recognition. In Advances in Speech Signal Processing, Eds. Furui, S. and Sondhi, M. Marcel Dekker, 701-738, 1992.

19. Sambur, M. R., Selection of acoustic features for speaker identification. IEEE Trans. Acous., Speech and Signal Proc. ASSP-23, 169-176, 1975.

20. Tosi, O., Oyer, H., Lashbrook, W., Pedrey, C. and Nash, W., Experiments on voice identification. JASA, 51, 2030-2043, 1972.

21. Vaissiere, J., Speech recognition: A tutorial. In Computer Speech Processing, Eds. Fallside, F. and Woods, W., Prentice Hall, 200-215, 1985.