SPRINGER
REFERENCE

Stan Z. Li
Editor

Anil K. Jain
Editorial Advisor

Encyclopedia of Biometrics

Springer

References

1. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE T. Acoust. Speech (ASSP-26) **26**(1), 43–49 (1978)
2. Booth, I., Barlow, M., Watson, B.: Enhancements to dtw and vq decision algorithms for speaker recognition **13**(3–4), 427–433 (1993)
3. Soong, F.K., Rosenberg, A.E., Rabiner, L.R., Juang, B.H.: A vector quantization approach to speaker recognition. Approach Speaker Recogn., **66**(2), 14–26 (1987)
4. Bimbot, F., Bonastre, J.F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska, D., Reynolds, D.A.: A tutorial on text-independent speaker verification. EURASIP Journal on Applied Signal Processing, Special issue on biometric signal processing (2004)
5. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted Gaussian mixture models. Digit. Signal Process, **10**(1–3), 19–41 (2000)
6. Wan, V.: Speaker Verification Using Support Vector Machines. Ph.D. thesis, University of Sheffield (2003)
7. Campbell, W., Campbell, J., Reynolds, D., Singer, E., Torres-Carrasquillo, P.: Support vector machines for speaker and language recognition. Comput. Speech Lang., **20**(2–3), 210–229 (2006)
8. Ben, M., Betser, M., Bimbot, F., Gravier, G.: Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted gmms. In: ICSLP (2004)
9. Campbell, W.M., Sturim, D., Reynolds, D.A.: Support vector machines using GMM supervectors for speaker verification. IEEE Signal Process. Lett. **13** (2006)
10. Scheffer, N., Bonastre, J.F.: A UBM-GMM driven discriminative approach for speaker verification. In: Odyssey (2006)
11. Rabiner, L., Juang, B.: Fundamentals of Speech Recognition. Prentice-Hall, Upper Saddle River, (1992)
12. Nordstrm, T., Melin, H., Lindberg, J.: comparative study of speaker verification systems using the polycost database. In: International Conference on Spoken Language Processing ICSLP (1992)
13. Tishby, N.: On the application of mixture ar hidden markov models to text-independent speaker recognition. pp. 563–570 (1991)
14. Reynolds, D., Carlson. B.: Text-dependent speaker verification using decoupled and integrated speaker and speech recognizers. In: EUROSPEECH in Madrid, ESCA (1995)

Speaker Model

Speaker model is a representation of the identity of a speaker obtained from a speech utterance of known origin. It can be generative or discriminative. Most popular generative speaker models are the Gaussian Mixture Models (GMM), which model the statistical distribution of speaker features with a mixture of Gaussians. Typical discriminative speaker models are based on the use of Support Vector Machines (SVM), where the speaker model is basically a separating hyperplane in a high-dimensional space. Once enrolled, speaker models may be compared to a set of features coming from an utterance of unknown origin, to give a similarity score.

► Speaker Features

Speaker Parameters

► Speaker Features

Speaker Recognition Engine

► Speaker Matching

Speaker Recognition, One to One

► Liveness Assurance in Voice Authentication

Speaker Recognition, Overview

Jean Hennebert
Department of Informatics, University of Fribourg, Fribourg, Switzerland
Institute of Business Information Systems HES-SO Valais, TechnoArk, Sierre, Switzerland

Synonyms

Voice recognition; Voice biometric

Definition

Speaker recognition is the task of recognizing people from their voices. Speaker recognition is based on the extraction and modeling of acoustic features of speech that can differentiate individuals. These features conveys two kinds of biometric information: physiological properties (anatomical configuration of the vocal apparatus) and behavioral traits (speaking style). Automatic speaker recognition technology declines into four major tasks, *speaker identification*, *speaker verification*, *speaker segmentation*, and *speaker tracking*. While these tasks are quite different for their potential applications, the underlying technologies are yet closely related.

Introduction

Speaking is the most natural mean of communication between humans. Driven by a great deal of potential applications in human-machine interaction, automated systems have been developed to automatically extract the different pieces of information conveyed in the speech signal (Fig. 1). Speech recognition systems attempt to transcribe the content of what is spoken. Language identification systems aim at discovering the language in use. Speaker recognition systems aim to discover information about the identity of the speaker.

Interestingly, speaker recognition is one of the few biometric approach which is not based on image processing. Speaker dependent features are actually indirectly measured from the speech signal which is 1-dimensional and temporal. Speaker recognition is a biometrics qualified as *performance-based* or *active* since the user has to cooperate to produce a sequence of sounds. This is also a major difference with other *passive* biometrics such as for fingerprints, iris, or face recognition systems where user cooperation is not requested.

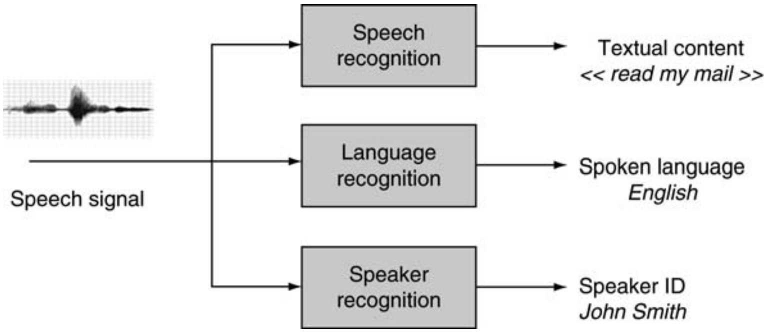
Speaker recognition technologies are often ranked as less accurate than other biometric technologies such as fingerprint or iris scan. However, there are two main factors that make voice a compelling biometric. First, there is a proliferation of automated telephony services for which speaker recognition can be directly applied. Telephone handsets are indeed available basically everywhere and provide the required sensors for the speech signal. Second, talking is a very natural gesture and it is often considered as lowly intrusive by users as no physical contact is requested. These two factors, added to the recent scientific progresses, made speaker recognition converge into a mature technology.

Speaker recognition finds applications in many different areas such as access control, transaction authentication, forensics, speech data management, and personalization. Commercial products offering voice biometric are available from different vendors. However, many technical and non-technical issues, discussed in the next sections, still remain open and are still subjects of intense research.

History of Speaker Recognition

Research and development on speaker recognition methods and techniques have now spanned more than five decades and it continues to be an active area [1].

In 1941, the laboratories of Bell Telephone in New Jersey produced a machine able to visualize spectrograph of voice signals. During the Second World War, the work on the spectrograph was classified as a military project. Acoustic scientists used it to attempt to



Speaker Recognition, Overview. Figure 1 The different speech tasks can be declined into speech recognition, language identification, and speaker recognition.

identify enemy voices from intercepted telephone and radio communications. In the 1950's and 1960's, so-called *Experts* testimony in forensic application started. These experts were claiming that spectrographs were a precise way to identify individuals, which is of course not true in most conditions. They associated the term "voiceprint" to spectrographs, as a direct analogy to fingerprint [2]. This *expert* ability to identify people on the basis of spectrographs was very much disputed in the field of forensic applications, for many years and even until now [3].

The introduction of the first computers and mini-computers in the 1960's and 1970's triggered the beginning of more thorough and applied research in speaker recognition [4]. More realistic access control applications were studied incorporating real-life constraints as the need to build systems with single-session enrolment. In the 1980's, speaker verification began to be applied in the telecom area. Other application issues were then uncovered, such as unwanted variabilities due to microphone and channel. More complex statistical modelling techniques were also introduced such as the Hidden Markov Models [5]. In the 1990's, common speaker verification databases were made available through the Linguistic Data Consortium (LDC). This was a major step that triggered more intensive collaborative research and common assessment. The National Institute of Standards and Technology (NIST) started to organize open evaluations of speaker verification systems in 1997.

In the present decade, the recent advances in computer performances and the proliferation of automated system to access information and services pulled speaker recognition systems out of the laboratories into robust commercialized products. Currently, the technology remains expensive and deployment still needs lots of customization according to the context of use. From a research point of view, new trends are also appearing. For example, the extraction of higher-level information such as word usage or pronunciation is studied more for applications and new systems are attempting to combine speaker verification with other modalities such as face [6, 7] or handwriting [8].

Speech Signal

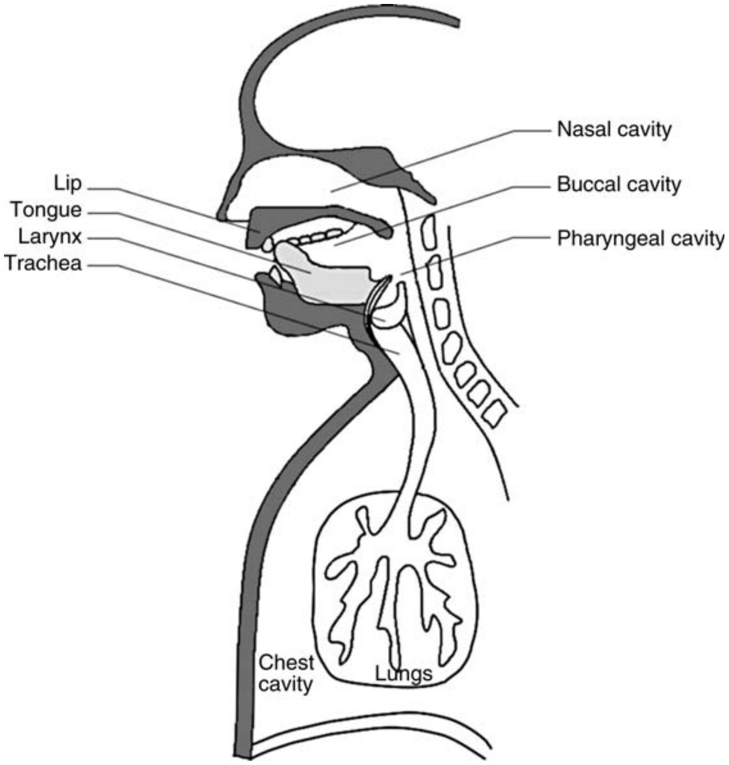
Speech production is the result of the execution of neuromuscular commands that expel air from the lungs, causes vocal cords to vibrate, or to stay steady and shape the tract through which the air is flowing out. As illustrated in Fig. 2, the *vocal apparatus* includes three cavities. The pharyngeal and buccal cavities form the *vocal tract*. The nasal cavity form the *nasal tract* that can be coupled to the vocal tract by a trap-door mechanism at the back of the mouth cavity. The vocal tract can be shaped in many different ways determined by the positions of the lips, tongue, jaw, and soft palate.

The *vocal cords* are located in the larynx and, when tensed, have the capacity to periodically open or close the larynx to produce the so-called *voiced sounds*. The air is hashed and pulsed in the vocal apparatus at a given frequency called the *pitch*. The sound then produced resonates according to the shapes of the different cavities. When the vocal cords are not vibrating, the air can freely pass through the larynx and two types of sounds are then possible: *unvoiced sounds* are produced when the air becomes turbulent at a point of constriction and *transient plosive sounds* are produced when the pressure is accumulated and abruptly released at a point of total closure in the vocal tract.

Roughly, the speech signal is a sequence of sounds that are produced by the different articulators changing positions over time [9]. The speech signal can then be characterized by a time-varying frequency content. Figure 3 shows an example of a voice sample. The signal is said to be slowly time varying or quasi-stationary because when examined over short time windows (Fig. 3-b), its characteristics are fairly stationary (5 − 100 msec) while over long periods (Fig. 3-a), the signal is non-stationary (>200 msec), reflecting the different speech sounds being spoken.

The speech signal conveys two kinds of information about the speaker's identity:

1. *Physiological properties*. The anatomical configuration of the vocal apparatus impacts on the production of the speech signal. Typically, dimensions of the nasal, oral, and pharyngeal cavities and the length of vocal cords influence the way phonemes are produced. From an analysis of the speech signal, Speaker recognition systems will indirectly capture some of these physiological properties characterizing the speaker.
2. *Behavioral traits*. Due to their personality type and parental influence, speakers produce speech with different phonemes rate, prosody, and coarticulation



Speaker Recognition, Overview. Figure 2 Schematic view of the human vocal apparatus. The vocal apparatus includes three cavities: the pharyngeal, buccal, and nasal cavities. These cavities form the vocal and nasal tract that can be shaped in many different ways determined by the positions of the lips, tongue, jaw, and soft palate.

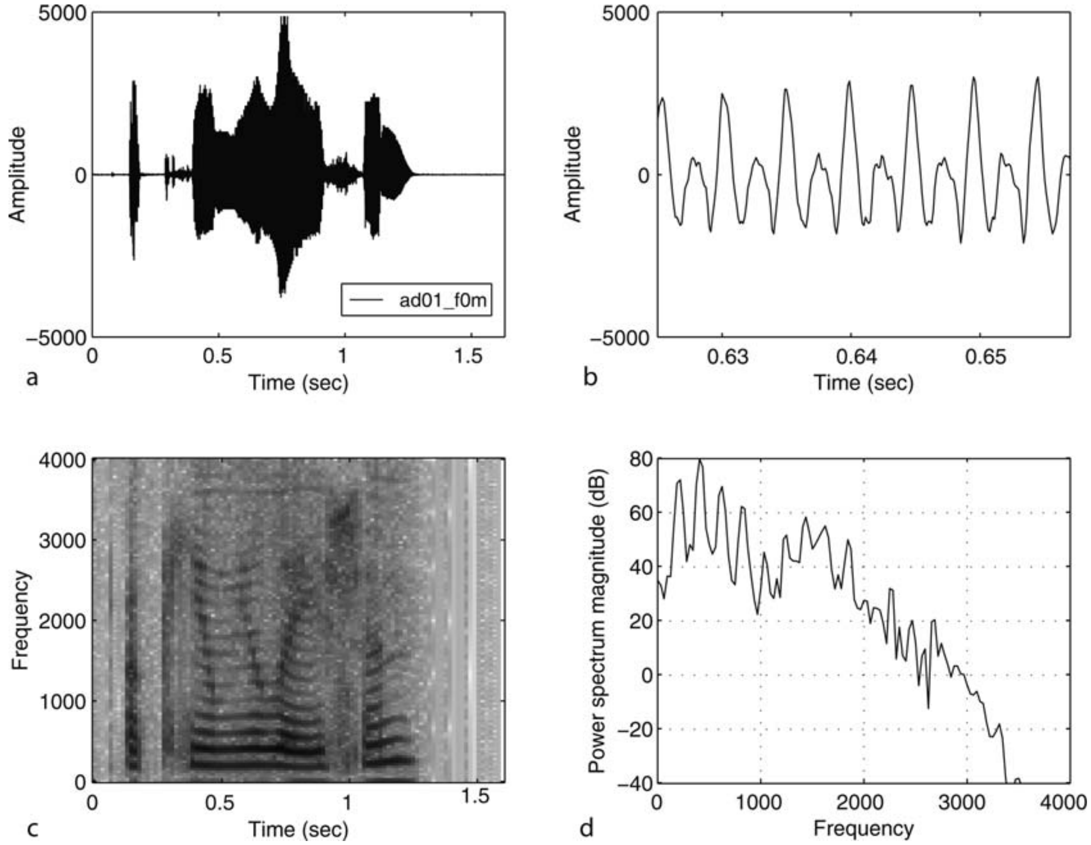
effects. Due to their education, socio-economic status, and environment background, speakers use different vocabulary, grammatical constructions, and diction. All these higher-level traits are of course specific to the speaker. Hesitation, filler sounds, and idiosyncrasies also give perceptual cues for speaker recognition.

Most of the speaker recognition systems are relying on low-level acoustic features that are linked to the physiological properties. Some behavioral traits such as prosody or phoneme duration are partly captured by some systems. Higher-level behavioral traits such as preferred vocabulary are usually not implicitly modeled by speaker recognition systems because they are difficult to extract and model. Typically, the system would need a large amount of enrolment data to determine the preferred vocabulary of a speaker, which is not reasonable for most of the commercial applications.

Intra-speaker variabilities are due to differences of the state of the speaker (emotional, health, . . .). Inter-speaker variabilities are due to physiological or behavioral differences between speakers. Automatic speaker recognition systems exploit inter-speaker variabilities to distinguish between speakers but are impaired by the intra-speaker variabilities which are, for the voice modality, numerous.

Feature Extraction and Modeling

In the case of the speech signal, the feature extractor will first have to deal with the long-term nonstationarity. For this reason, the speech signal is usually cut into frames of about 10-30 msec and feature extraction is performed on each piece of the waveform. Secondly, the feature extraction algorithm has to cope with the short-term redundancy so that a reduced and relevant acoustic information is extracted. For this purpose, the representation of the waveform is generally swapped from the temporal domain to the frequency domain, in which the short-term temporal periodicity is represented by higher energy values at the frequency



Speaker Recognition, Overview. Figure 3 Speech signal of the word *accumulation*: (**a**) waveform, (**b**) partial waveform, (**c**) narrow-band spectrogram of (**a**), (**d**) power spectrum magnitude of (**b**).

corresponding to the period. Thirdly, feature extraction should smooth out possible degradations incurred by the signal when transmitted on the communication channel. For example, in the case of telephone speech, the limited bandwidth and the channel variability will need some special treatment. Finally, feature extraction should map the speech representation into a form which is compatible with the statistical classification tools in the remainder of the processing chain.

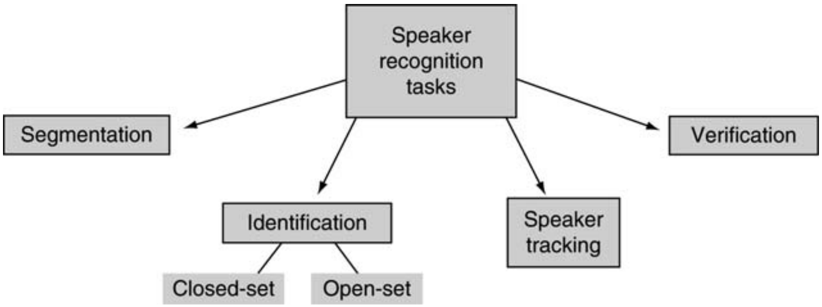
Usual feature extraction techniques are the so-called *linear predictive coding (LPC) cepstral* analysis or the *mel-frequency cepstral* analysis. These algorithms are widely used in the field of speech processing [9, 10]. The output of the feature extraction module is a temporal sequence of acoustic vectors $X = \{x_1, x_2, \ldots, x_N\}$ of length N with each vector x_n having a constant dimension D. The sequence X is then input into the pattern classification module.

There are many different ways reported in the scientific literature to build speaker models: vector quantization, second order statistical methods, Gaussian Mixtures Model (GMM), Artificial Neural Network (ANN), Hidden Markov Model (HMM), Support Vector Machines (SVM), etc. One of the most widely used is GMM modeling. By nature, GMMs are versatile as they can approximate any probability density function given a sufficient number of mixtures. With GMMs, the probability density function $p(x_n|M_{client})$ or *likelihood* of a D-dimensional feature vector x_n given the model of the client M_{client}, is estimated as a weighted sum of multivariate gaussian densities (e.g., [11]).

Speaker Recognition Tasks and Applications

Automatic speaker recognition can be declined into four tasks (Fig. 4).

Speaker identification attempts to answer the question "Whose voice is this?" In the case of large speaker



Speaker Recognition, Overview. Figure 4 From left to right, the different speaker recognition tasks can be loosely classified from the most difficult to the less difficult ones. The tasks of verification and identification are the major ones considering the potential commercial applications.

sets, it can be a difficult task where chances are more to find speakers with similar voice characteristics. The identification task is said to be *closed-set* if it is sure that the unknown voice comes from the set of enrolled speaker. By adding a "none-of-the-speaker" option, the task becomes an *open-set* identification. Speaker identification is mainly applied in surveillance domains and, apart from this, it has a rather small number of commercial applications. *Speaker verification* (Also known as *speaker detection* or *speaker authentication* task.) attempts to answer the question "Is this the voice of Mr Smith?" In other words, a candidate speaker claims an identity and the system must accept or reject this claim. Speaker verification has a lot of potential commercial applications thanks to the growing number of automated telephony services. When multiple speakers are involved, these tasks can be extended to *speaker tracking* (when a given user is speaking) and *speaker segmentation* (blind clustering of a multi-speaker record).

Speaker recognition systems can also be classified according to the type of text that the user utters to get authenticated. One can distinguish between ► *text-dependent*, ► *text-prompted*, and ► *text-independent* systems. These categories are generally used to classify speaker verification tasks. To some extent, they can also apply to the task of identification.

- *Text-dependent systems*. These systems use the same piece of text for the enrolment and for the subsequent authentication sessions. Recognition performances of text-dependent systems are usually good. Indeed, as the same sequence of sounds is produced from session to session, the characteristics extracted from the speech signal are more stable. Text-dependency also allows to use finer modeling techniques capable to capture information about sequence of sounds. A major drawback of text-dependent systems lies in the replay attacks that can be performed easily with a simple device playing back a pre-recorded voice sample of the user. The term *password-based* is used to qualify text-dependent systems where the piece of text is kept short and is not supposed to be known to other users. There are *system selected text/password* where an a priori fixed phrase is composed by the system and associated to the user (e.g., pin codes) and *user selected text/password* where the user can freely decide on the content of the text.
- *Text-prompted systems*. Here the sequence of words that need to be said is not known in advance by the user. Instead, the system prompts the user to utter a randomly chosen sequence of words. A text-prompted system actually works in two steps. First, the system performs speech recognition to check that the user has actually said the expected sequence of words. If the speech recognition succeeds, then the verification takes place. This *challenge-response* strategy achieves a good level of security by preventing replay attacks.
- *Text-independent*. In this case, there is no constraint on the text spoken by the user. The advantages are the same as for the text-prompted approach: no password needs to be remembered and the system can incrementally ask for more data to reach a given level of confidence. The main drawback lies here in the vulnerability against replay attacks since any recording of the user's voice can be used to break into the system.

Speaker recognition finds applications in many different areas such as telephony transaction authentication, access control, speech data management, and forensics. It is in the telephony services that speaker recognition finds the largest deal of applications as the technology can be directly applied without the need to install any sensors.

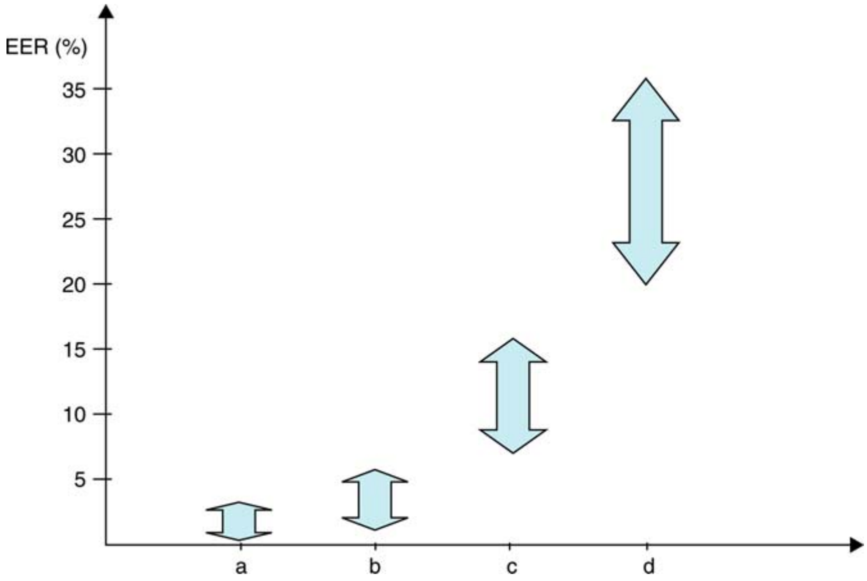
- *Telephony authentication for transactions*. Speaker recognition is the only biometric that can be directly applied to the automated telephony services (Interactive Voice Response - IVR systems). Speaker recognition technology can be used to secure the access to reserved telephony services or to authenticate the user while doing automated transactions. Banks and telecommunication companies are the main potential clients for such systems. As many factors impact on the performances of speaker recognition in telephony environment, it is often used as a complement to other existing authentication procedures. Most of the implementations are using a text-prompted procedure to avoid pre-recording attacks and to facilitate the interaction with a dialog where the user just needs to repeat what the system is prompting. A less known but interesting example of speaker verification application in telephony is also the home incarceration and parole/probation monitoring.
- *Access control*. Speaker verification can be used for physical access control in combination with the usual mechanisms (key or badge) to improve security at relatively low cost. Applications such as voice-actuated door locks for home or ignition switch for automobile are already commercialized. Authorized activation of computers, mobile phones, or PDA is also an area for potential applications. Such applications are often based on text-dependent procedures using single passwords.
- *Speech data management and personalization*. Speaker tracking can be used to organize the information in audio documents by answering the questions: who and when a given speaker has been talking? Typical target applications are in the movie and media industry with speaker indexing and automatic speaker change detection for automatic subtitling. Automatic annotation of meeting recordings and intelligent voice mail could also benefit from this technology. In the area of personalization, applications to recognize broad speaker characteristics such as gender or age can be used to personalize advertisements or services.
- *Forensic speaker recognition*. Some criminal cases have recordings of lawbreakers voice and, speaker verification technologies can help the investigator in directing the investigation. On the other hand, there is a general acceptation in the scientific community on the fact that a verification match obtained with an automatic system or even with a so-called voiceprint expert, should not be used as a proof of guilt or innocence [3].

Performances and Influencing Factors

Figure 5 summarizes typical ranges of Equal Error Rate (EER) performances for four categories of speaker verification systems [12]. The range of performances is globally extremely large, going from 0.1 to 30% across the systems. Text-dependent applications using high quality speech signals can have very low EER typically ranging from 0.1 to 2%. Such performances are obtained with multi-session enrolment of several minutes and test data of several seconds acquired in the same condition as for the enrolment. Pin-based text-dependent applications running on the telephony channel will typically show performances ranging from 2 to 5%. Text-independent applications based on telephony quality, recorded during conversations over multiple handsets and using several minutes of multi-session enrolment data and a dozen of seconds for the test data, will show EER ranging from 7 to 15%. Finally, text-independent applications based on very noisy radio data will show performances ranging from 20 to 35%.

Summary

Speaker recognition is often ranked as providing medium accuracy in comparison to other biometrics. This is due to three main factors. First, there are the inherent and numerous intra-speaker variabilities of the speech signal (emotional state, health condition, age). Second, the inter-speaker variabilities are



Speaker Recognition, Overview. Figure 5 Typical performances of speaker verification systems. The arrows define ranges of Equal Error Rates for four different types of applications. Applications of type (**a**) are text-dependent based on high quality speech signals. Applications of type (**b**) are text-dependent based on telephony speech quality, typically a pin-based application. Applications of type (**c**) are text-independent on telephony speech quality recorded during conversations. Applications of type (**d**) are text-independent based on very noisy radio.

relatively weak, especially within family members. Finally, the speech signal is often exposed to all sort of environmental noise and distortions due to the communication channel. These varying acquisition conditions are captured by the speech template which becomes biased. To smooth out these variabilities, lengthy or repeated enrollment sessions are often performed, but this is generally at the expense of usability.

Speaker recognition remains however a compelling biometrics. First, talking is considered a very natural gesture and user acceptance is generally high. Furthermore no physical contact is requested to record the biometric sample and the rate of failure to enroll is also very low. Finally, the technology cost of ownership is pretty low. For computer-based applications, simple sound cards and microphones are available at low-cost. For telephony applications, there is no need for special acquisition devices as any handset can be used from basically anywhere.

Speaker recognition technology has made tremendous progress over the past 20 years and finds new applications in many different areas such as telephony authentication, access control, law enforcement, speech data management, and personalization.

Related Entries

- ▶ Biometrics, Overview
- ▶ Speaker Feature
- ▶ Session Effects on Speaker Modeling
- ▶ Speech Analysis
- ▶ Speech Production

References

1. Furui, S.: 50 years of progress in speech and speaker recognition. In: Proceedings of SPECOM, pp. 1–9 (2005)
2. Kersta, L.: Voiceprint Identification. Nature **196**, 1253–1257 (1962)
3. Boe, L.J.: Forensic voice identification in France. Speech Commun. **31**, 205–224 (2000)
4. Atal, B.S.: Automatic recognition of speakers from their voices. Proc. IEEE **64**, 460–475 (1976)
5. Naik, J.M., Netsch, L.P., Doddington, G.R.: Speaker verification over long distance telephone lines. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Glasgow, Scotland pp. 524–527 (1989)
6. Jain, A., Ross, A., Prebhakar, S.: An introduction to biometric recognition. IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics **14**(1) (2004)

7. Fauve, B., Bredin, H., Karam, W., Verdet, F., Mayoue, A., Chollet, G., Hennebert, J., Lewis, R., Mason, J., Mokbel, C., Petrovska., D.: Some results from the biosecure talking face evaluation campaign. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Las Vegas, USA (2008)
8. Humm, A., Hennebert, J., Ingold, R.: Spoken signature for user authentication. SPIE J. Electron. Imaging, Special Section on Biometrics: ASUI **17**(1) (2008)
9. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice Hall (1993)
10. Picone, J.: Signal modeling techniques in speech recognition. Proc. IEEE **81**(9), 1214–1247 (1993)
11. Reynolds, D.: Automatic speaker recognition using gaussian mixture speaker models. Linc. Lab. J. **8**(2), 173–191 (1995)
12. Reynolds, D.: An overview of automatic speaker recognition technology. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 4, pp. 4072–4075 (2002)

Speaker Recognition, Standardization

JUDITH MARKOWITZ
Consultants, Chicago, IL, USA

Synonyms

Speaker authentication; Speaker biometrics; Speaker identification and verification, SIV; Voice authentication; Voice recognition

Definition

The term "speaker recognition" (SR) refers to a group of technologies that use information extracted from a person's speech to perform biometric operations such as speaker identification and verification (SIV). Standards for SR are designed to support the development of applications that can work with technology from different vendors (application programming interface standards), the sharing of SR data (data interchange standards), the transmission of data in real time (distributed speaker recognition standards), and the management of data resources in distributed environments (process-control protocol standards).

Introduction

SR technologies stand at the juncture between ► speech-processing and biometrics. They belong in speech processing, because they extract and analyze data from the ► stream of speech. They belong in biometrics, because the data that are extracted describe a physical or behavioral characteristic of the speaker and because they use that information to make decisions regarding the speaker, usually determining the identity of the speaker and verifying a claim of identity. Some SR technologies perform other speaker-related functions, such as placing the speaker into a category, such as female or male (► speaker classification); determining whether the speaker has changed (speaker change); assessing the speaker's level of stress or emotion (emotion detection, voice stress analysis); tracking a specific voice in a multispeaker communication (speaker/voice tracking); separating interleaved and overlapping voices from each other (► speaker separation); and determining whether the speaker is lying or telling the truth (voice lie detection).

Standards for SR come from both speech processing and from biometrics. They fall into several categories:

1. Application programming interface (API) standards,
2. Sharing of stored SR data (data interchange),
3. Transmission of data in real time (distributed speaker recognition) and
4. Management of data resources in distributed environments (process-control protocols).

Application Programming Interface (API) Standards – Early Work

API standards eliminate the need for programmers to learn a new set of programming functions for each SR product. They accomplish this by establishing a standard set of functions that can be used to develop applications using any standards-compliant SR technology.

The bulk of the work on SR standards has been directed toward the development of standard APIs. Most of these standards have been crafted by speech-processing industry consortia or standards bodies and are extensions of existing standards for ► speech recognition.