About Speaker Recognition Techology

Gerik Alexander von Graevenitz

Bergdata Biometrics GmbH, Bonn, Germany

gerik @ graevenitz.de

Overview about Biometrics

Biometrics refers to the automatic identification of a living person based on physiological or behavioural characteristics. There are many types of biometric technologies on the market: face recognition, fingerprint recognition, finger geometry, hand geometry, iris recognition, vein recognition, voice and signature recognition.

The method of biometric identification is preferred over traditional methods involving passwords and PIN numbers for various reasons:

The person to be identified is required to be physically present at the point-of-identification. The identification based on biometric techniques obviates the need to remember a password or carry a token or a smartcard.

With the rapid increase in use of PINs and passwords occurring as a result of the information technology revolution, it is necessary to restrict access to sensitive/personal data. By replacing PINs and passwords, biometric techniques are more convenient in relation to the user and can potentially prevent unauthorised access to or fraudulent use of ATMs, Time & Attendance Systems, cellular phones, smart cards, desktop PCs,

Workstations, and computer networks. PINs and passwords may be forgotten, and token based methods of identification, like passports, driver's licenses and insurance cards, may be forgotten, stolen, or lost.

Various types of biometric systems are being used for real-time identification; the most popular are based on face recognition and fingerprint matching. Furthermore, there are other biometric systems that utilise iris and retinal scan, speech, face, and hand geometry.

Voice Recognition

Speech contains information about the identity of the speaker. A speech signal includes also the language this is spoken, the presence and type of speech pathologies, the physical and emotional state of the speaker. Often, humans are able to extract the identity information when the speech comes from a speaker they are acquainted with.

LAWRENCE KERSTA at the Bell Labs made the first major step from speaker verification by humans towards speaker verifications by computers in the early 1960s where he introduced the term voiceprint for a spectrogram, which was generated by a complicated electro-mechanical device. The voiceprint was matched with a verification algorithm that was based on visual comparison.

The recording of the human voice for speaker recognition requires a human to say something. In other words the human has to show some of his/her speaking behavior. Therefore, voice recognition fits within the category of behavioral biometrics.

A speech signal is a very complex function of the speaker and his environment that can be captured easily with a standard microphone. In contradiction to a physical biometric technology such as fingerprint, in speaker recognition are not fixed, no static and no physical characteristics. In speaker recognition there are only information depending on an act.

The state of-the-art approach to automatic speaker verification (denoted as ASV) is to build a stochastic model of a speaker, based on speaker characteristics extracted from the available amount of training speech.

In speaker recognition we differ between low-level and high-level information. High level-information are values like a dialect, an accent, the talking style and the subject manner of context. These features are currently only recognitized and analyzed by humans. As low-level are denoted the information like pitch period, rhythm, tone, spectral magnitude, frequencies, and bandwiths of an individual's voice. These features are used by speaker recognition systems.

Voice verification works with a microphone or with a regular telephone handset, although performance increases with higher quality capture devices. The hardware costs are very low, because today nearly every PC includes a microphone or it can be easily connected one. However voice recognition has got its problems with persons who are husky or mimic another voice. If this happens the user may not be recognized by the system. Additionally, the likelihood of recognition decreases with poor-quality microphones and if there is background noise. Voice verification will be a complementary technique for e.g. finger-scan technology as many people see finger recognition technology as a higher authentication

form. In general voice authentication has got a high EER, therefore it is in general not used for identification. The speech is variant in time, therefore adaptive templates or methods are necessary.

Intraspeaker variance versus Interspeaker variance

The variation of features caused by different speakers is called interspeaker variance. The interspeaker variance is caused by different vocal characteristics of individuals and provides useful information for distinguishing different speakers. Another kind of variation – intraspeaker variation occurs when a speaker pronounces the same word or sentence but cannot repeat the utterance in exactly the same way from trial to trail.

The intraspeaker variation includes the different speaking rate, the emotional state of the speaker and the speaking environment. The intraspeaker variation is the main factor that causes the performance degradation of speaker recognition systems. Therefore, it is desirable to select the parameters that show lower intraspeaker but high interspeaker variability. In many speaker recognition applications, it is possible to reduce the intraspeaker variability by requiring the user to pronounce the test sentence that contains the same text or vocabularies as the training sentences. This is the case of text-dependent speaker recognition methods.

Text-dependent vs. text-independent speaker recognition.

Speaker recognition systems are classified as text-dependent (fixed-text) and text-independent (free-text). The text-dependent systems require a user to repronounce some specified utterances, usually containing the same text as the training data. There is no such constraint in text-independent systems. In the text-dependent system, the knowledge of knowing words or word sequence can be exploited to improve the performance.

There are two main reasons for wanting a speaker verification system to prompt the client with a new password phrase for each new test occasion: (a) The user does not have to remember a fixed password and (b) the system can not easily be defeated with the replaying of recordings of the user´s speech.

There are a few methods that are used for speaker verification. The text-dependent speaker recognition methods can be classified into DTW (dynamic time warping) or HMM (Hidden Markov Model) based methods.

Text-independent speaker verification has been an active area of research for a long time because performance degradation due to mismatched conditions has been a significant barrier for deployment of speaker recognition technologies.

How it works.

There are a few methods that are used for speaker verification. The text-dependent speaker recognition methods can be classified into DTW (dynamic time warping) or HMM (Hidden Markov Model) based methods.

The DTW-methods are using instantaneous and transitional cepstra. In 1963, Bogert et al. published a paper with the title “The Quefrency Analysis of Time Series for Echoes”. They defined a new signal processing technique where they defined an extensive vocabulary interchanging letters like the word spectrum in cepstrum. For the computation of the cepstrum usually a Fast Fourier Transformation is used.

Since 1975 the Hidden Markov Modeling (denoted as HMM) is a technique that has become popular in speech recognition research, named by the Russian mathematician A.A. Markov. With HMM-based methods, the statistical variation of spectral features is measured.

Examples for the text-independent speaker recognition methods are: the average-spectrum-based method, the VQ-based methods and the multivariate auto-regression (MAR) model.

The average-spectrum-based method is using a weighted cepstral distance measure, where the phoneme effects in speech spectra are removed by averaging the spectra.

With the VQ-based method a set of short-term training feature vectors of a speaker can be used directly to represent the essential characteristics of that speaker. However, such a direct representation is impractical when the number of training vectors is large, since the memory and amount of computation required become prohibitively large. Therefore, attempts have been made to find efficient ways of compressing the training data using vector quantization (VQ) techniques.

Montacie et alii applied a multivariate auto-regression (MAR) model to the time series of cepstral vectors to characterize speakers with receiving quite good results.

Anyways, the text-independent speaker verification has been an active area of research for a long time because performance degradation due to mismatched conditions has been a significant barrier for deployment of speaker recognition technologies.

Sources of Verification Errors

There are a few sources of verification errors that may occur:

- Misspoken or misread prompted phrases
- Extreme emotional states (e.g. stress or duress)
- The attitude how the speech is said is another than with the enrollment
- Time varying (intra- or intersession) microphone placement
- Poor or inconsistent room acoustics (e.g. multipath and noise)
- Channel mismatch (e.g. using different microphones for enrollment

and verification)

- Different pronunciation speed during the verification compared with the training data.
- Sickness (e.g. head colds can alter the vocal tract)
- Aging (the vocal tract can drift away from models with age)
- Women have a quite higher FRR, because the spectral of the voice is smaller.

Faking a voice verification system requires a very high quality recorder, which is not easy to find on the market. Normal voice recorders that are on the market do not record the complete spectrum of the voice that is necessary to fake the system. The quality loss of the voice recording system must be very low, too. With the most voice verification systems imitating a voice from one human by another human does not lead to success.

The mentioned sources of verification errors lead to the result that actually it is quite complex to do an identification with voice verification. Therefore the voice verification systems are used for verification in most cases in combination with a PIN or a chipcard to identify the user in a database.

Applications

The application for speaker verification systems are:

- Time and Attendance Systems
- Access Control Systems

- Telephone-Banking/Broking
- Biometric Login to telephone aided shopping systems
- Information and Reservation Services
- Security control for confidential information
- Forensic purposes

Conclusion

The advantage of a voice verification system is the very cheap hardware that is needed – in most computers a soundcard and a microphone is implemented. It use very easy to use and to implement with applications for the telecommunication.

Voice recognition has got a few disadvantages, too. On the one hand the human voice is not invariant in time therefore the biometric template must be adapted during progressing time. The human voice is also variable through temporal variations of the voice, caused by a cold, hoarseness, stress, emotional different states or puberty vocal change. On the other hand voice recognition systems have got a higher EER compared to fingerprint recognition systems, because the human voice is not as “unique” as fingerprints. For the computation of the Fast Fourier Transformation the systems needs to have a co-processor and more processing power than e.g. for fingerprint matching. Therefore speaker verification systems are not suitable for mobile applications / battery powered systems in the current state.