Please repeat: "*my voice is my password*" From the basics to real-life implementations of speaker verification technologies

Jean Hennebert

Université de Fribourg, Boulevard de Pérolles 90, 1700 Fribourg, Switzerland
`jean.hennebert`

Abstract. Speaker verification finds applications in many different areas such as access control, transaction authentication, law enforcement, speech data management and personalization. As for other biometric technologies the prime motivation of speaker recognition is to achieve a more usable and reliable personal identification than by using artifacts such as keys, badges, magnetic cards or memorized passwords. Speaker verification technologies are often ranked as less accurate than other biometric technologies such as iris scan or fingerprints. 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. Second, talking is a very natural gesture, often considered as lowly intrusive by users as no physical contact is requested. These two factors, added to the recent scientific progresses, made voice biometric converge into a mature technology.

1 Introduction

Speaking is the most natural mean of communication between humans. Driven by a great deal of potential applications in human-machine interaction, systems have been developed to automatically extract the different pieces of information conveyed in the speech signal. There are three major tasks. In *speech recognition* tasks, the automatic system aims at discovering the sequence of words forming the spoken message. In *language recognition* tasks, the system attempts to identify the language used in a given piece of speech signal. Finally, *speaker recognition* systems aim to discover information about the identity of the speaker. Speaker recognition finds applications in many different areas such as access control, transaction authentication, law enforcement, speech data management and personalization. As for other biometric technologies the prime motivation of speaker recognition is to achieve a more usable and reliable personal identification than by using the classical methods based, for example, on keys, badges, magnetic cards or memorized passwords.

Interestingly, speaker recognition is one of the few biometric approach which is not based on image processing. Speaker recognition systems are often said to be *performance-based* since the user has to produce a sequence of sound. This is also a major difference with other *passive* biometrics for which the cooperation of the authenticated person is not requested, such as for fingerprints, iris or face recognition systems. Speaker recognition technologies are often ranked as less accurate than other biometric technologies such as finger print 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, often considered as lowly intrusive by users as no physical contact is requested. These two factors, added to the recent scientific progresses, made voice biometric converge into a mature technology. Commercial products offering voice biometric are now available from different vendors. However, many technical and non-technical issues, discussed later in this chapter, still remain open and need to be tackled. Also, the technology remains expensive and deployment still needs lots of customization according to the context of use. A list of speaker recognition vendors can be found in [1]. From a research point of view, new trends are also appearing. For example, the extraction of higher-levels information such as word usage or pronunciation is more and more studied. In other example, we see new systems attempting to combine speaker verification with other modalities such as face, lips movements or handwriting.

1.1 Speaker Recognition Tasks

As illustrated on Figure 1, automatic *speaker recognition* technology declines into four major tasks, *speaker identification*, *speaker verification*, *speaker segmentation* and *speaker tracking*[1]. While these tasks are quite different by their potential applications, the underlying technologies are yet closely related.

The *speaker identification* task attempts to answer the question "Whose voice is this?". An identification task aims at associating an unknown voice with one from a set of N known, enrolled speakers. It can be a difficult task in the case of large speaker sets where chances are higher 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.

The *speaker verification* task[2] attempts to answer the question "Is this the voice of Mr Smith?". In other words, a candidate speaker claims an identity and

[1] Another speaker authentication technology, known as *verbal information verification* [6] aims at authenticating speakers by verifying the content of the spoken utterance against the user's personal profile. This methodology is more related to a password-based authentication method than to a real biometric technology and is not further described here.

[2] Also known as *speaker detection* or *speaker authentication* task.

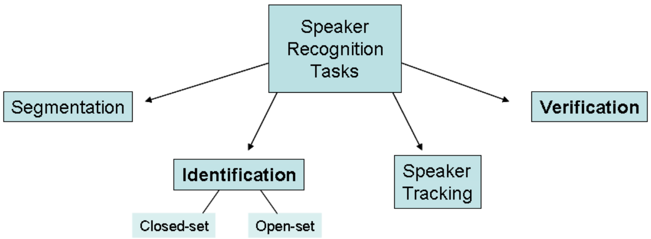

Fig. 1. The different speaker recognition tasks. From left to right, the tasks are loosely classified from the most difficult to the less difficult ones. The tasks of verification and identification are the major ones considering the potential applications.

the system must accept or reject this claim. Speaker verification is a much simpler classification problem with only two categories. The first category represents the *true* speaker[3] and the second category represents the other speakers[4]. Speaker verification has a large deal of commercial applications thanks to the growing number of automated telephony services. For example, banks are rolling out phonebanking services to let their clients perform transactions for which the security could be successfully reinforced through the use of speaker verification.

There exist also two extended tasks that are considered in the area of multi-speaker recognition. The *speaker tracking* task aims at determining if a target user speaks in a multi-speaker record and at finding the speech stretches corresponding to this speaker. The *segmentation* task is probably the most difficult task as it aims at blindly clustering a multi-speaker record. No assumptions about the number of speaker and no a priori training data is available for this segmentation task. These tasks find applications in the area of automatic indexation of meeting records or broadcasting programme.

1.2 Pro and cons of Speaker Recognition

Speaker recognition, as all biometrics, has strengths and limitations. Systems based on speaker recognition are therefore not the answer to every context of use. It is then crucial to understand the pros and cons of speaker recognition before proceeding with purchase and deployment of this technology.

- **Pro: High user acceptance**. Analyzing the speech signal produced by a speaker is usually considered as lowly intrusive. Talking is indeed a very natural gesture for human and no physical contact is requested by the device to record the biometric sample. Sometimes, the speaker does not even realize that an authentication is occurring. Excepted for impaired person, the rate of failure to enroll is also very low.

[3] Also known as *client* or *target* speaker.

[4] Also known as *impostor* speakers

- **Pro: Good security against intentional attacks**. When a higher level of security is needed against lawbreakers intentionally trying to claim someone else's identity, speaker recognition can be implemented with a *text-prompted* methodology allowing challenge-response strategies (see section 3.2). With such systems, simple pre-recording of the voice sample is not enough to break the authentication procedure.

- **Pro: Low technology cost**. 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 phone handset can be used.

- **Pro: Remote network of sensors already available**. The fixed and mobile telephone networks provide application access point from almost anywhere, without the need of any dedicated hardware. The internet also provides access points for any connected computers equipped with a microphone.

- **Cons: Medium accuracy**. Speaker recognition is recognized to provide medium accuracy in comparison to other biometrics. This is due to three main factors. First, there are the inherent *intra-speaker* variabilities of the speech signal which are induced by the emotional state of the user, by his health condition and finally by his age. For example, a simple cold catched by a registered user can increase the chance of him being rejected. Second, the *inter-speaker* variabilities are relatively weak, i.e. chances to find people with similar speaking characteristics are not as low as for finger print or iris scan. Users of the same family or twins are, for example, more difficult to distinguish by speaker recognition systems since their vocal apparatus characteristics may be very similar. Finally, the speech signal can be 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. When considering telephony-based applications for example, mismatches of microphones as well as telephony channel are known to impair speaker recognition performances. More is said about factors impacting accuracy in Section 4.

- **Cons: Repeated enrollment sessions**. The user is asked to record a given quantity of speech during one or more enrolment sessions. Few enrolment data will yield most of the time to biased voice templates. The length and the amount of enrolment sessions will then condition the quality of the templates. Now, a tradeoff must be reached because for the user, enrolment should be as short as possible while for the system, enrolment should be long enough to cover most of the variabilities. Therefore, acceptable levels of accuracy often require to perform repeated enrolment sessions, sometimes spaced in time. Some systems circumvent this weakness by implementing hidden incremental enrolment to make the voice template more robust while the speaker is using the system.

2 Speech signal

2.1 Speech production

The speech signal 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. Roughly, the speech signal is a sequence of sounds that are produced by the different articulators changing positions over time [8]. Figure 2-a 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 (Figure 2-b), its characteristics are fairly stationary $(5 - 100$ msec$)$ while over long periods (Figure 2-a), the signal is non-stationary $(> 200$ msec$)$, reflecting the different speech sounds being spoken. The *phonemes* are the linguistically distinct speech sounds in a given language. Phonemes are produced with variable speed and frequencies. 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.

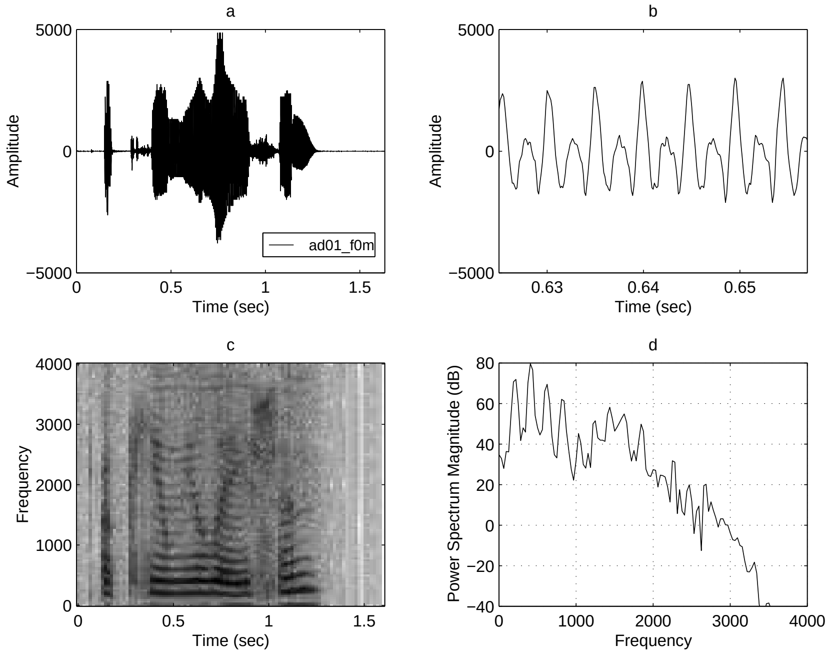

Fig. 2. Speech signal of the word *accumulation*: (a) waveform, (b) partial waveform, (c) narrow-band spectrogram of (a), (d) power spectrum magnitude of (b).

2.2 Speaker specific features

Roughly, the speech signal itself 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. Physiological characteristics such as the nasal, oral and pharyngeal cavity dimensions or the vocal cords length influence the way phonemes are produced. 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 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 are also giving perceptual cues identifying somehow the speaker.

Most of state-of-the-art speaker recognition systems are relying on low-level acoustic features closely linked to the physiological properties. Some behavioral traits such as prosody or phoneme duration are partly captured by these systems. Higher-level behavioral traits such as preferred vocabulary are usually not implicitly modelled by speaker verification 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 of course not reasonable for most of the commercial applications.

3 Speaker Recognition Systems

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

3.1 Text-Dependent

Text-dependent 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 modelling techniques capable to capture more detailed 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[5].

[5] To increase somehow the security, the system can keep track of the previous user attempts (or a signature of it) in order to reject attempts made by recordings.

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 by other users (difficult to do when the password is said aloud). From a very practical point of view, the deployment of password-based systems necessitates the creation of a service that needs to deal with renewal of passwords when they are lost or forgotten. The management of this service, automated or not, can be a real burden.

We can further classify text-dependent and password-based systems into two categories.

- *System selected text/password :* An a priori fixed phrase is composed by the system and associated to the user. The phrase is usually built from words taken in a limited lexicon. Systems based on *Personal Identification Numbers* (PIN) are a typical example. These systems generate different sequences of digits for each user. Such systems are not user-convenient because the user has to remember a given text. However, this procedure has one major advantage for speaker verification tasks. It allows to perform the identity claim and the verification step at the same time, using the same speech sample that is first recognized and then verified.
- *User selected text/password :* In this case, the user can freely decide on the content of the text. This is a more user friendly procedure. However, such systems are potentially more difficult to develop than in the previous case. First, the modelling technique can not take advantage of the a priori knowledge of the text, leading potentially to weaker performances. Second, the implementation needs an extra layer of automatic supervision to discard too short or too long passwords.

Text-dependent systems are mostly used in speaker verification for access control applications.

3.2 Text-Prompted

With text-prompted systems, 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-prompted strategy can actually be seen as a liveness test of the user[6].

From an ergonomic point of view, no text or password needs to be remembered, which is nowadays appreciated by the users. On the other side, the procedure request a dialog with the user who has to listen and repeat a given utterance.

[6] Intentional attacks are still possible by building a voice modification or synthesis engine to reproduce the voice of the target user. Nevertheless, this requires specific knowledge and a pretty large amount of speech signal from the target user.

Performances of text-prompted systems are usually very good. Indeed, the modelling technique can take advantage of the a priori knowledge of the text. Also, the procedure can be set up to prompt the user to provide more data when the confidence in the decision is not high enough. Text-prompted systems are mainly used for access control applications. They are for example well-suited to secure the access of voice-activated telephony services as an automated dialog with the user is already running.

3.3 Text-Independent

In the case of text-independent systems, there is no constraint on the text spoken by the user. As the amount of enrolment data is limited, the different realizations of the sounds are not all captured by the system. For this reason, the resulting accuracy is often lower than for text-dependent systems. Text-independent systems are mostly used for speaker identification tasks such as telephony surveillance, automatic indexation of broadcast shows or meeting recordings. They are also used in most of the forensic applications where the speaker is usually not cooperative to say a given text. They can also be used for access control applications. 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.

4 Performances

4.1 Typical performances of verification systems

We report here typical speaker verification performances for some broad categories of systems. The reader should be aware that there are many factors that influence the performances. These factors are varying very much with the enrolment and testing scenarios. It is therefore quite difficult to define clear-cut frontiers of performances and the figures provided here should be interpreted with caution.

A speaker recognition system outputs a score, noted here R_c that is proportional to the likelihood that the input speech signal has been produced by the speaker model under test. More specifically, for a verification task, the system will decide to accept the claimed identity if this score R_c is above a given decision threshold T and to reject it in the other case. The system can then make two types of errors. It can decide to falsely accept an impostor or it can decide to falsely reject a true client. Results of speaker verification systems are classically measured in terms of impostor *False Acceptation* (FA) and client *False Rejection* (FR) error rates that vary as a function of the decision threshold T.

Benchmarks are usually performed by running the system on a test set including genuine and impostor accesses. Operating points (FA, FR) can then be plot on a (x, y) figure with T as parameter. The special point where $FA = FR$ is called the Equal Error Rates (EER) and is classically used as a summary indicator of performances.

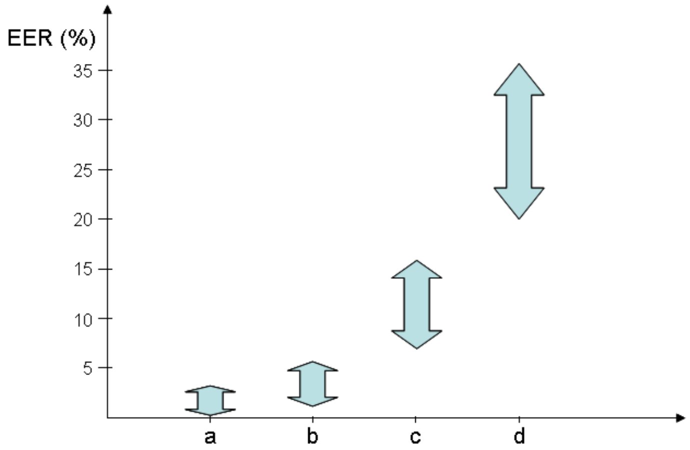

Fig. 3. Typical performances of speaker verification systems. The arrows define ranges of Equal Error Rates for 4 different types of applications. Applications of type (**a**) are text-dependent based on high quality speech signals. Applications of type (**b**) are text-dependent based on telephony speech 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.

Figure 3 summarizes typical ranges of EER performances for 4 categories of speaker verification systems [9]. 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%.

4.2 Influencing factors

As general rules, increasing the speech quality, increasing the control of the acquisition, increasing the amount of training and testing data, and limiting the vocabulary will enhance the performances. We give more details here on the different factors that influence the performance of speaker recognition systems:

- **User cooperation**: Commercial speaker recognition systems are generally used in applications in which the user wishes to be recognized and is therefore cooperative. Text-dependent or text-prompted systems are usually deployed in such cases. The user will generally be careful to speak clearly and intelligibly to the system in a non-adverse environment. Such commercial systems will generally reach a good level of performance. In some other situations, the user will not be inclined to cooperate or will not know that he is under a recognition process (background verification). For example, in the case of conversational speech, the voice will potentially reflect all kind of emotional states, increasing the intra-speaker variabilities of the speech signal characteristics and leading also to less accuracy for the verification system.
- **Recording conditions**: The quality of the recording is maybe the most influential factors on the accuracy of speaker recognition systems. Environmental noise, background conversation, channel variabilities are known to decrease the quality of the speech signal and then the accuracy of the system [7]. The problem comes also partly from the speaker modelling technique which captures some of the characteristics of the acquisition environment. Mismatched conditions between enrolment and tests usually lead to a drop of the performance. When possible, normalization techniques are applied to make speaker modelling independent from the conditions of acquisition but a full de-correlation of the speaker and environment characteristics is usually not possible. A classic problem for telephony based applications is the mismatch due to the handset microphones which can have different response curve or due to the channel variations when changing from a land-line to a mobile telephone.
- **Amount of enrolment and test data**: The duration of the enrolment session has also a large impact on the accuracy. In short, the more sessions and the longer the enrolment, the better the accuracy. Using one short enrolment sessions (few dozen of seconds) is often imposed by the application designer for usability reasons, i.e. the user does not want to spend a long time to enrol. Unique and short enrolment sessions unfortunately will not cover all the variabilities of the speaker voice and therefore, the template may reveal biased or incomplete.
- **Intra-speaker variabilities**: *Long-term*: In general, the effect of aging is sufficiently slow to allow to use the same model for a long time without

[7] This is especially true when degraded conditions occur during the enrolment session, leading to a low-quality model of the speaker which is then used as the reference for the next verification attempts of the user.

re-enrolling. Long-term evolutions due to aging can be tracked with incremental enrolment techniques. *Short-term*: Unexpected transitional states of the user can also degrade the accuracy. The effects of drugs, sickness or strong emotional events may affect the performance of the system.

As explained above, the accuracy of speaker recognition systems depends strongly on the scenario of the target application. Furthermore, many parameters are influencing the performances. It is therefore very difficult to compare the effective performance of different systems. The scientific community has now recognized the need to assess systems and algorithms on common tasks as a primary activity for driving the technology forward. In this direction, the Speech Group of the National Institute of Standards and Technology (NIST) has now been coordinating evaluations of speaker verification technologies since 1997, with an increasing success over the years.

5 Potential Applications

There are many potential applications to speaker recognition. Currently, the most known commercial applications are telephony based where speaker recognition is the only biometric that can be directly applied. As discussed in section 4, speaker recognition technology is not absolutely reliable as many factors impact on the performances. Therefore, the tendency is to use speaker recognition as a complement to other existing authentication procedure to diminish the level of frauds.

5.1 Telephony authentication for transactions

Many companies such as banks or telecom providers have been rolling out for years automatic telephony systems to allow their customers performing basic operations from any telephone. These systems are running on Interactive Voice Response (IVR) telephony platforms that are parts or extensions of the classical Public Branch Exchange (PBX) telephony system. The IVR dialogs are either touch-tone using the Dual Tone Multiple Frequency (DTMF) signals or voice activated with speech recognition. Speaker recognition technology is sometimes used in these IVR systems. As lawbreakers may try to take the place of the user once the authentication step has been successfully passed, the verification system can also be advantageously used for monitoring that the speaker did not change during the call.

In Phonebanking systems, speaker verification technology can be used as a replacement or as an addition to the traditional pin-based authentication procedure[8]. Most of the implementations are using a text-prompted procedure to

[8] Some banks still offer non-automated operator based services. In this case, the authentication is often performed using verbal queries where the user is asked for passwords or personal information. This procedure is weaker as the information can easily be reproduced by lawbreakers.

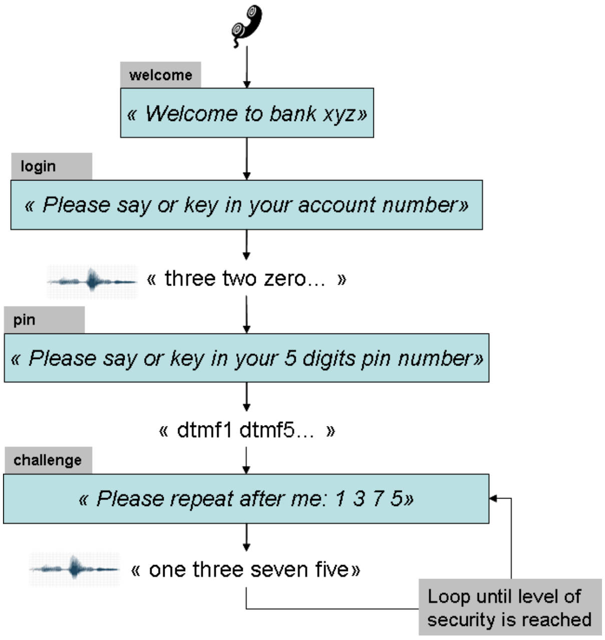
welcome
« Welcome to bank xyz»
login
« Please say or key in your account number»
« three two zero... »
pin
« Please say or key in your 5 digits pin number»
« dtmf1 dtmf5... »
challenge
« Please repeat after me: 1 3 7 5»
« one three seven five»
Loop until level of security is reached

Fig. 4. Typical dialog flow of a phonebanking application.

avoid pre-recording attacks and to facilitate the interaction with a dialog where the user just need to repeat what the system is prompting (see Section 3.2). Figure 4 illustrates a typical dialog flow of a text-prompted phonebanking system. The text-prompted part occurs in the *challenge* dialog state where the user is asked to repeat, for example, a sequence of 4 digits. The system often loops in this state until enough speech samples are accumulated to reach a given level of confidence in the decision. Typically, three challenges of 4 digits are enough to reach good performances. Of course, all the samples of speech signal can actually be used as material for the speaker verification module to take a decision, such as, for example, the one of the login state in Figure 4. If the system is still not confident, the call may be routed to a human operator who will handle the call.

Telecom providers are also large providers of automated IVR telephony services. They have realized that frauds can be reduced with speaker verification systems, for example in the case of calling card applications. As the transaction is not as risky as in the case of phonebanking, the calling card number and eventually the pin code said by the user can be directly used to perform authentication, without the need for a challenge-response procedure.

Finally, all telephony services using credit cards transactions could benefit from speaker verification technology. There are, for example, ticketing or teleshopping services which are known to be a source of unauthorized use of credit cards.

5.2 Access control

Physical access control systems could use speaker verification 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 mobile phones or PDA is also an area for potential applications. Such applications are often based on text-dependent procedures using single passwords (see Section 3.1). Access control to computer networks can also be performed using speaker verification, in addition to the usual password procedure. However in this case, the implementation is less obvious. This is firstly due to the fact that computer microphones have very diverse quality leading to mismatched signal characteristics. Secondly, users may not have the privilege or habit to access the different microphone settings. Finally and maybe more importantly, users are generally not used to talk to computers. An interesting case is the one of some large companies that are using speaker verification to automate computer password reset procedures. When an employee forgets a password, a secured procedure needs to be put in place to renew the password. Such procedures can cost very much if performed manually and are a source of potential security breach if simply automated. The use of speaker verification may allow to automate fully the procedure while keeping a fairly good level of security.

5.3 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. Meeting recordings could also benefit from this technology. Such applications have very different characteristics than the one for access control. In particular, there is potentially a large amount of training material to build the speaker model and the processing time is usually not limited. More advanced modelling techniques can then be used to enhance the accuracy of the recognition. However, the tracking task is potentially more difficult as speakers may speak simultaneously.

Speaker identification can be used to label automatically voice mails and allow for advanced browsing or specific actions, such as forwarding to the secretary or to a colleague. Intelligent answering machines could also identify the caller and provide personal reply. In the area of personalization, applications to recognize broad speaker characteristics such as gender or age can be used to personalize advertisements or services. Finally, another prospective area is games where speaker recognition is applied to recognize the owner or to make interactivity more personal.

5.4 Law enforcement and forensic

A less known but interesting example of speaker verification application is the home incarceration and parole/probation monitoring. Several experiments with home incarceration have already started around the world, motivated by costly overcrowded jails and by the increasing awareness of the negative effects of imprisonment. Home incarceration implies a procedure to verify, in an non-intrusive way, that the parolees are indeed staying home. This can be done by calling the parolees at random times of the day. A speaker verification system can then provide an efficient way to guarantee the identity of the one who is answering the call. Another similar application is prison call monitoring. Some inmates have restricted permissions to perform outbound calls from the prison. In this case, a speaker verification system can be used to verify the identity of the inmate prior to allow the outbound call.

Criminal cases have sometimes access to recordings of the voice of lawbreakers. Speaker verification can here be used as a help in directing an investigation. Interestingly, the scientific community agreed 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 [5]. This movement has been initiated as an acknowledgement of the limits of the technology and of the many different factors that may impact on the verification reliability. Another area close to forensic applications is the automated surveillance of telephone calls.

6 NIST speaker recognition evaluation

In the 1990's, the scientific community realized the importance of comparing algorithms using common speaker verification databases. The National Institute of Standards and Technology (NIST) also triggered more intensive collaborative research and common assessment by organizing open evaluations of speaker verification systems on common tasks [2]. The initial evaluations from 1996 to 2000 were targeting tasks where the quantities of data to train and test the systems were limited. This was corresponding to realistic conditions, close to commercial applications of speaker verification. From 2001, the tendency was clearly to evaluate the algorithms on more training and testing data, i.e. targeting tasks probably related to surveillance applications. Table 1 summarizes the evolution of the NIST evaluations from 1996 to 2006.

Years	Train (sec)		Test (sec)		Main analyzed factors
	min	max	min	max	
1996-1998	60	180	3	30	Handset variation and test duration
1999-2000	120	120	1-3	60	Additional tasks on speaker tracking
2001-2003	120	3600	15	45	More data to train, cellular data
2004-2006	300	3600	10	300	More data to test, multilingual data

Table 1. Evolution of the NIST evaluations through years. The minimum and maximum available quantities of speech data to train and test systems are reported, as well as the main analyzed factor.

7 Conclusion

Speaker recognition technology has made tremendous progress over the past 20 years and finds now applications in many different areas such as telephony authentication, access control, law enforcement, speech data management and personalization. Often ranked as less accurate than other biometric technologies, speaker recognition remains a compelling biometric for two main reasons. First, there is a proliferation of automated telephony services for which speaker recognition is the only biometric that can be directly applied. Telephone handsets are available from everywhere and provide the required sensors for the speech signal. Second, talking is considered as a very natural gesture and is lowly intrusive as no physical contact with the sensor is requested. These two factors, added to the recent scientific progresses, made voice biometric converge into a mature technology.

Further readings We recommend [4] and [9] as good tutorials and overviews of speaker recognition technology. For further information on the algorithms, we refer to [7] for the feature extraction part and to [10] and [3] for state-of-the-art classification algorithms.

References

1. http://www.biometrics.org/.
2. http://www.nist.gov/speech/.
3. Younès Bennani and Patrick Gallinari. Connectionist approaches for automatic speaker recognition. In *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pages 95–102, Martigny, Switzerland, April 1994.
4. F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacrtaz, and D. Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 4:430–451, 2004.
5. Louis-Jean Boe. Forensic voice identification in france. *International Conference Speech and Computer*, (31):205–224, 2000.
6. Chin-Hui Lee. A unified statistical hypothesis testing approach to speaker verification and verbal information verification. In *Speech Technology in the Public Telephone Network, Where Are We Today ?*, pages 63–72, Rhodes, 1997.
7. Joseph Picone. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9):1214–1247, September 1993.
8. Lauwrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition.* Prentice Hall, 1993.
9. Douglas Reynolds. An overview of automatic speaker recognition technology. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 4, pages 4072–4075, 2002.
10. Douglas Reynolds, Thomas Quatieri, and Robert Dunn. Speaker verification using adapted gaussian mixture models. 10:19–41, 2000.