Journal of InterDisciplinary Research Applied to Medicine
JDREAM. Journal of interDisciplinary REsearch Applied to Medicine
JDREAM (2018), v. 2 i. 1, 5-16
ISSN 2532-7518
DOI 10.1285/i25327518v2i1p5
http://siba-ese.unisalento.it, © 2018 Università del Salento

Voice Biometrics Technologies and Applications for Healthcare: an overview.

Francesco Sigona[1]

[1]Centre for Interdisciplinary Research on Language (CRIL), University of Salento, Italy
Department of Engineering, University of Sannio, Italy

Abstract
Our medical information is perhaps considered the most personal and valuable of all. On the other hand, the rapid increase in the digitization of the healthcare industry nowadays is giving the patient identification and electronic medical record maintenance a key role to improve patient safety, protect against medical identity theft, enhance patient satisfaction, not to mention financial performances. Regulatory pressures and compliance guidelines naturally require health care providers to handle patient information with the highest degree of privacy and care. In this scenario, innovative healthcare organizations are increasingly using biometric technologies to support smoother and safer patient access and information sharing to improve security. Biometric identifiers are measurable unique characteristics of an individual physiological (such as fingerprints, iris, retina, DNA) or behavioural (such as typing rhythm, voice, and gait) traits that can be used to positively identify a person. Among them, voice biometrics is actually a convenient and secure method of authenticating a speaker's identity using just a few words, which has many advantages over other forms of biometric identification especially in the healthcare sector. After briefly discussing the impact of biometric technologies from a general point of view, the present paper focuses on voice biometrics, to give a technical understanding of how it works according to the current state of art and the forthcoming progress, and how it can be successfully integrated in the healthcare systems.

Keywords: voice biometrics, speaker recognition, patients identification, staff authentication.

Introduction

Among the many challenges facing healthcare today is the need of costs containment while improving the quality of care. To overcome these challenges, it seems increasingly necessary to efficiently ensure the proper identification of all the healthcare constituencies (staff, providers, payers and patients). The failure to correctly identify patients opens the doors to frauds and, even more seriously, to patient identity (ID) theft. The identity security is even more important in the healthcare sector than in others, because other than costly medical errors, identity theft can prevent legitimate people to access critical medical services, with consequences that can become dramatic. Moreover, unlike other sectors such as financial, frauds can easily go undetected for a long time. Healthcare frauds can be done, for example, by providing false information when applying for programs or services, forging prescriptions or selling prescription drugs. Also, medical records have a very high transactional value in black markets compared, for instance, to stolen credit information, because of the huge amount of data about a particular victim that is available through these records. Moreover, patient records are sensitive and their safety is a regulatory requirement. Although statistics do not always agree about the costs of healthcare frauds, some authoritative estimates ranged around 6.19% of global healthcare expenditure, equivalent to €350 billion (PKF Lillteljohn LLP and University of Portsmouth, 2015).

The most important reason why such kind of attacks have been able to occur is the vulnerability of account protection and authentication methods, and this is why the healthcare industry is being increasingly attracted by the biometric technologies. This is confirmed also by the most recent market researches, that estimate that the global healthcare biometrics market will grow at an impressive CAGR of 22.9% to surpass $13.9 million in revenue in 2025

(http://www.credenceresearch.com/report/healthcare-biometrics-market), with North America and Europe together capturing over 70% of the global healthcare biometrics market.
Biometrics in the healthcare industry can be used for applications in doctors' offices, hospitals, or for use in monitoring patients, to implement logical access control to centralized archives of digitized patients' data, physical access control to buildings and hospital wards, workforce management, patient record storage, authenticate medical and social support personnel. Voice biometrics is a mature enough technology to provide the required levels of security, versatility, cost effectiveness, preference by users, which only are required to speak to get automatically recognized.
This paper is organized as follows. Section 2 introduces some essentials about biometric recognition principles and definitions. Sections 3 and 4 gives a more technical overview of the voice biometrics concepts and the techniques used up to the current state of the art. In Section 5 the main advantages of voice biometrics over other kinds of biometric recognition in healthcare are discussed. Finally, conclusions are presented in Section 6.

Biometrics: some fundamentals.
The terms "Biometrics" and "Biometry", from the Greek root of βίος (bíos, i.e. "life") and μέτρο (mètron, i.e. "measure"), literally mean the "measurement of life", and refer to a methodological process that confirms its classification as a science. Biometrics, therefore, is a systematization of applicable knowledge to experimentally controllable hypotheses (Preite, 2016). Since the studies started in 1890 by Francis Galton (1822-1911) and later by Karl Pearson (1857-1936), the terms "Biometrics" and "Biometry", have been used traditionally in reference to mathematical methods to address data analysis problems in the biological sciences, indeed they are also referred to as "biological statistics" or "biostatistics".
On the other hand, recently, the term "Biometrics" is increasingly used to refer to the (automatic) identification and authentication of an individual by means of one or more measurable unique traits (Figure 1). Such biometric identifiers can be divided in two categories: "Anatomic-Physiological", and "Behavioural". The first category includes face traits and thermography, fingerprints, hand geometry, palm veins, iris, retina and DNA recognition, while typing rhythm, voice and gait come under behavioural biometrics.
With respect to traditional approaches that identify a person on the basis of a key or a token, physical (something an individual has) or intangible (something an individual knows, e.g. a password, a Personal Identification Number, PIN), in the biometric approach the individual itself becomes the key (something an individual is). Moreover, biometrics can be used in combination with those traditional approaches, leading to three-factors security level (Figure 2).
Generally, biometric schemes fall into two categories: "Authentication" and "Identification". In the authentication, the individual makes a claim of identity and the biometric recognizer has to verify if that claim is true or not, comparing the input features with those of the claimed identity (which was stored within an early enrollment stage). Authentication is therefore also referred to as "Verification", and has a binary output: accepted or rejected. In the identification approach there is no identity claim and the recognizer has to compare the input biometric features of the unknown individual against those of each individual in a set of pre-enrolled ones, producing a score for each comparison, and returning the identity of the enrolled person corresponding to the best score. If the system is allowed to reject the identification request when the highest score is still very low compared to a certain threshold, the identification is named "open-set"; on the contrary, if the system must return in any case the best matching identity, the identification is named "closed-set".

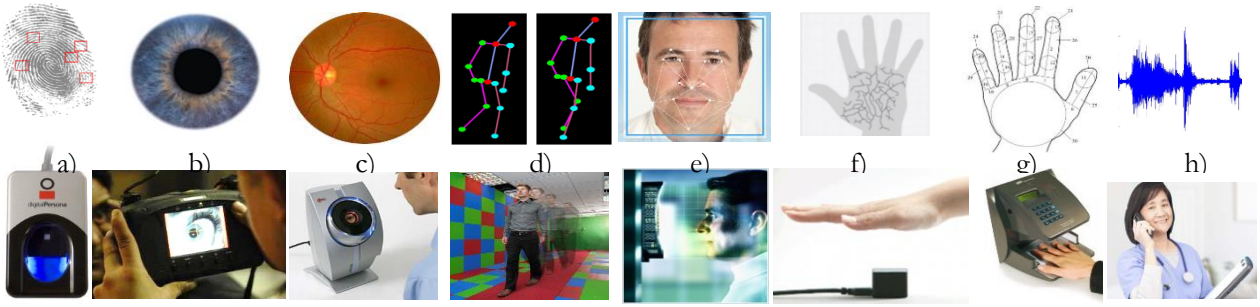

Figure 1. Some biometric identifiers and related devices. a) Fingerprint recognition looks for the unique patterns of ridges and valleys that are present in an individual's fingerprint, that is acquired by a specific hardware scanner, often in combination of temperature and pulse to improve the authority of the image; b) Iris recognition methodologies rely on image processing techniques to identify patterns in iris images, that need to be acquired by a scanning device; c) Retina recognition relies on the property of the blood vessels to absorb light more readily than the surrounding tissue; a retinal scanner is able to produce the appropriate lighting and acquiring the resulting pattern; d) Gait biometrics identifies and recognises people by the way they walk; individuals may be required to provide "gait" samples by walking through a tunnel where sensors are able to capture the relevant features; e) Face recognition algorithms typically identify facial features by extracting landmarks, or features, from an image of the subject's face; f) Palm vein recognition is generally based on the capability of deoxidized haemoglobin flowing through the veins to absorb near-infrared light, provided by a palm scanner., making them appear as dark pattern; g) Hand geometry recognition is based on measuring and recording the length, width, thickness, and surface area of an individual's hand while guided on a plate, by means of a specific device; h) voice biometrics is discussed more in detail in the present work.

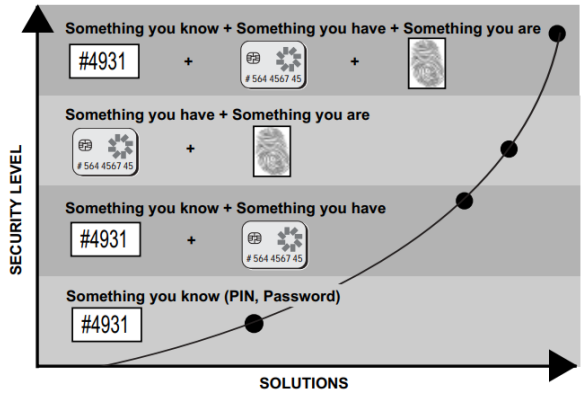

Figure 2: combining biometrics with traditional approaches improves the security level of authentication (Smart Card Alliance 2002).

As one can already guess by the above statements, biometric approaches are "Probabilistic", in the sense that the outcome is always intrinsically provided together with an estimate of the decision error. On the contrary, identification approaches such as the password-based one, in which the system can provide a sure match of all of the characters and case of the password, are named "Authoritative".

Performances of biometric approaches can therefore be measured on the basis of the extent of the errors they make while providing the outcome. Considering for instance a biometric authentication system, two kind of errors can be done: a "False Accept" (or "Miss"), when the system authenticates an impostor, and a "False Reject" when the system refuse a true identity claim. Accordingly, the system is designed to achieve an average "False Reject Rate" (FRR, also named type I error) and an average "False Accept Rate" (FAR, also named type II error), as a result of a trade-off between the two error types. Many other metrics can be defined on the basis of FAR and FRR, such as the "Equal Error Rate" (EER), which is the defined as the operating point where FAR equals the FRR, corresponding to an optimally chosen value of some parameter, which is typically a threshold for a some kind of similarity scoring function (see Figure 3).

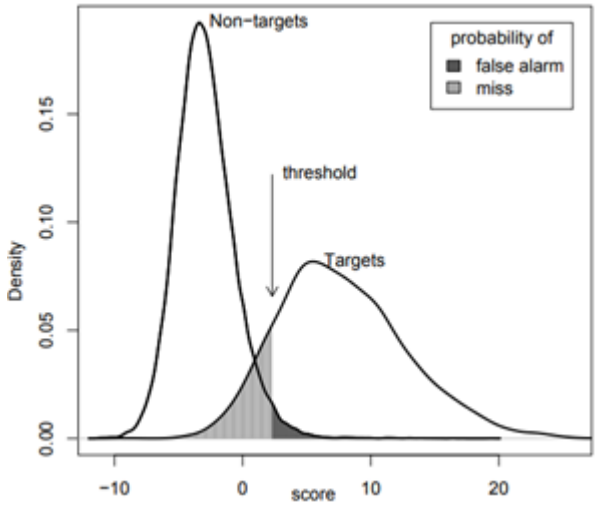

Figure 3: The score distributions for non-target (left) and target (right) trials. The grey areas left and right of the threshold represent P_{MISS} (which is equivalent to FRR) and P_{FA} (equivalent to FAR) respectively (van Leeuwen and Brümmer 2007).

On the other hand, the optimal operating point can also be chosen considering unequal costs for FA and FR errors (e.g. when higher FRR can be tolerated with respect to FAR), trying to minimize the following Detection Cost Function (DCF) introduced by the NIST (the well-known American National Institute of Standards and Technologies) over years:

$$DCF(\tau) = C_{MISS}P_{MISS}(\tau)P_{Target} + C_{FA}P_{FA}(\tau)(1$$

where P_{Target} is the prior probability of a specific target speaker, C_{MISS} and C_{FA} are fixed costs, that depend on the specific application, for the False Reject (i.e. Miss) and False Accept events, and P_{MISS} and P_{FA} are the probability of False Reject and False Accept events both as function of the score threshold τ. Both FAR and FRR depends dramatically on the chosen biometric identifiers and on the actual features that can be extracted from them. Provided that, in general, in biometric applications, no feature can assume always the same value over time, or, at least, can be extracted without some kind of superimposed noise, the key point is that the best features are those that, at the input of the biometric system, exhibit the smaller variability within the same individual, the greater variability across different individuals and the best robustness against noise. This is extremely important mainly for biometric identifiers such as the human voice, which is intrinsically characterized by a high degree of variability and, due to its versatility, can be used in many different scenarios and use cases in different environmental conditions.

Speaker recognition: the voice as biometric identifier.

Many dictionaries and textbooks use the term "Voice" and "Speech" interchangeably. However, from a scientific perspective, the two terms have different meanings. Voice (or vocalization) is the sound produced by the vibration of the vocal folds, in the larynx (or voice box), as the result of the pressure of the airflow produced by the lungs. Speech is an information-rich signal produced by a series of complex movements that, by means of frequency, amplitude and time modulation, change the basic tone created by voice into specific, decodable sounds (phonemes), that actually can be voiced or unvoiced (i.e. with or without vibration of the vocal cords), to convey information about words, identity, accent, expression, speaking style, emotion and the state of the speaker[1].

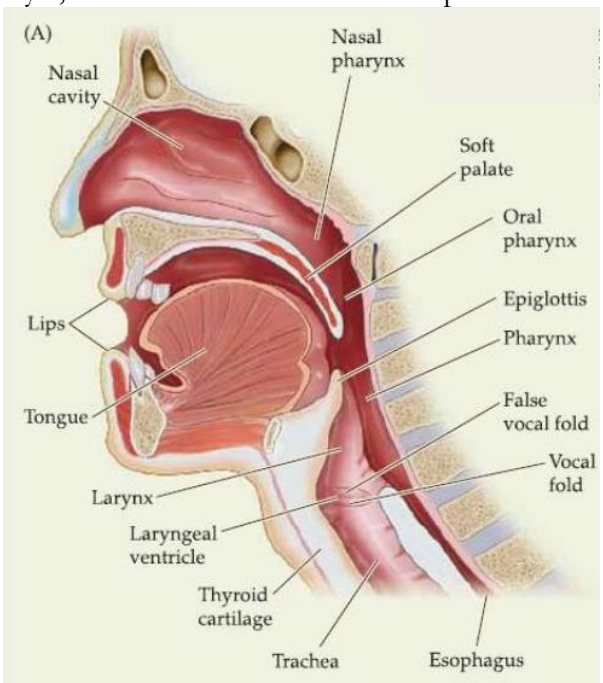

Figure 4: The peripheral phonation system (Purves, 2012).

1For sake of completeness, we report also a definition of "Language", as "the expression of human communication through which knowledge, belief, and behaviour can be experienced, explained, and shared. This sharing is based on systematic, conventionally used signs, sounds, gestures, or marks that convey understood meanings within a group or community" (National Institute on Deafness and Other Communication Disorders, https://www.nidcd.nih.gov/).

The distinction between Voice and Speech is also relevant to distinguish "Voice recognition" from "Speech recognition" applications. Voice recognition, Speaker recognition (or "Identification", "Authentication", "Verification") are all about the recognition of the speaker identity, i.e. the aim of Voice biometrics, while "Speech recognition" is largely used to address the task to recognize the spoken phonemes, words and sentences, i.e. the content of the speech. Strictly speaking, the recognition of what is said is not the field of voice biometrics, but as will be explained later in the present work, it can be successful combined with speaker recognition to increase the security of the voice-based recognition system, preventing impostors to be acknowledged thanks to fake pre-recorded utterances.

Speaker recognition is a complex operation that always start with the capture and processing of the speech signal, therefore, for this purpose, "Voice" and "Speech" can continue to be used interchangeably.

Speaker recognition is not a very recent concept, but, along with other biometric identification methodologies, it has become more accessible and affordable in recent times. Based on the pioneering studies started in 1960 by Gunnar Fant at the KTH Royal Institute of Technology (KTH, Swedish: Kungliga Tekniska högskolan) in Stockholm, Sweden, about the physiological components of acoustic speech production, and later expanded also by Joseph Perkell, in 1976 Texas Instruments built a first speaker recognition prototype system that was tested by the U.S. Air Force and the MITRE Corporation. Since then, a constant increasing interest has grown for the biometry of speech, and many progresses have been made in parallel with the development of technologies in the field of electronics, software engineering and digital signal processing and analysis.

The task of infer the identity of an individual by his/her voice is a challenging one, also because of the large intrinsic variability of the speech signal (other than the many factors depicted in Figure 5). However, this variability is associated with the ability to convey a huge variety and quantity of information, and it is what makes the voice the most natural form of human communication.

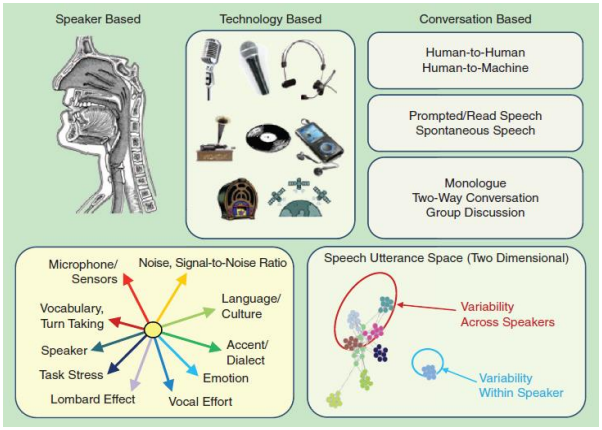

Figure 5: Source of variability in Speaker Recognition (Hansen and Hasan 2015).

In order the determine the identity of a speaker given one or more utterance, not the whole speech signal is considered, but only a set of measurements or "features". As the speech production is influenced by a given speaker in many different ways, features can be extracted at different levels as depicted in Figure 6. Other than maximizing the variability among different individuals and minimizing the variability within a given person, the ideal speaker features for biometric applications should also be difficult to mimic, be robust against the change in the speaker's health or long-term variations in voice, occur frequently and naturally in speech, be robust against noises and distortions.

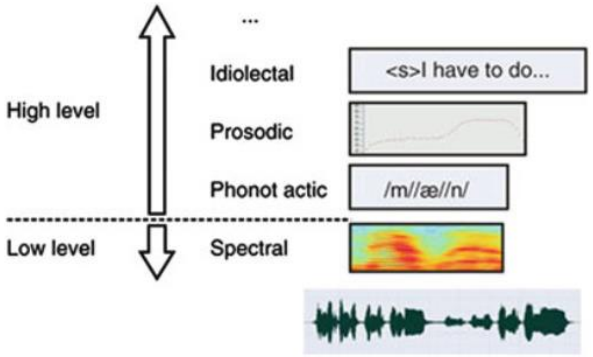

Figure 6: identity levels in the speech signal (Ramos et al. 2015).

Despite several studies shows the significance of suprasegmental (high level) features in speaker recognition systems (Demenko 2000, Yegnanarayana et al. 2001 and 2005), the low-level short-term spectral features are the most used, also because they reflect, more or less, the physical characteristics of the vocal tract. The concept of short-term spectral features can be easily depicted considering for example the Mel-Frequency Cepstral Coefficients (MFCCs,

Davis and Mermelstein, 1980), widely adopted features that have been proved to provide quite low error rates for clean signal conditions. Basically, the MFCCs are based on a linear cosine transform of a log power spectrum on a non-linear mel scale of frequency.

The whole utterance is divided in segments of duration of approximately 20-30 ms (hence the expression "short-term") with a certain amount of overlap (Figure 7). As depicted in Figure 8, for each frame, a power spectrum estimate is computed by means of Discrete Fourier Transform (DFT) and several other processing steps occur in order to achieve a bunch of numerical coefficients, which are the actual features. In the case of MFCCs, the power spectrum is input to a filter bank whose bands are equally spaced on the mel scale (see Figure 9), to better approximates the human auditory system's response.

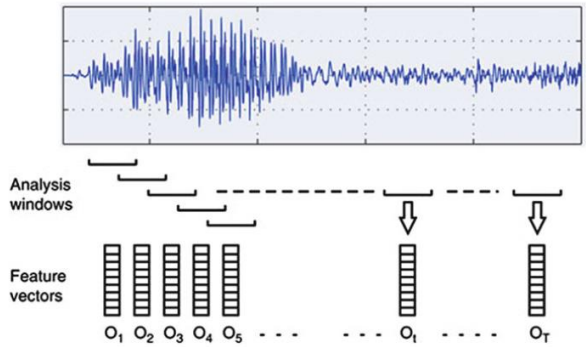

Figure 7: Overlapping framing of the acoustic waveform and frame-by-frame feature extractions. (Ramos et al. 2015).

Feature extraction outputs a temporal sequence of acoustic vectors $X = \{o_1, o_2, \ldots o_N\}$ of length N with each vector o_n having a constant dimension D.

Many other "flavours" of short-term spectral features have been proposed in order to increase robustness against noise and other disturbance affecting the speech signal, such as the Gammatone Frequency Cepstral Coefficients, GFCCs (Chanwoo et al. 2012), the Power Normalized Cepstral Coefficients, PNCCs (Al-Kaltakchi et al. 2016), and many others. Motivated by the fact that the neural responses are robust against noise, a research group has also proposed "neural" features as extracted from the responses of a physiologically-based computational model of the peripheral level of the auditory system (Islam et al. 2016).

The robustness against noise and channel mismatch is one of the true challenge for speaker recognition systems, mainly for those applications where the quality of the captured speech signal cannot be guaranteed. Even the environmental noise cannot always guaranteed to be a stationary process, because of the large variability of noise sources (reverberations, mechanical noises, even undesired voices of other speakers – the "cocktail party" effect). The speaker features may also get affected by the so called "channel mismatch", when the same individual could present himself to the recognizer using many devices – PC, smartphone, etc.- and transmission channels – mobile or landline telephone, data connections with different encodings and capacity), depending on the use cases.

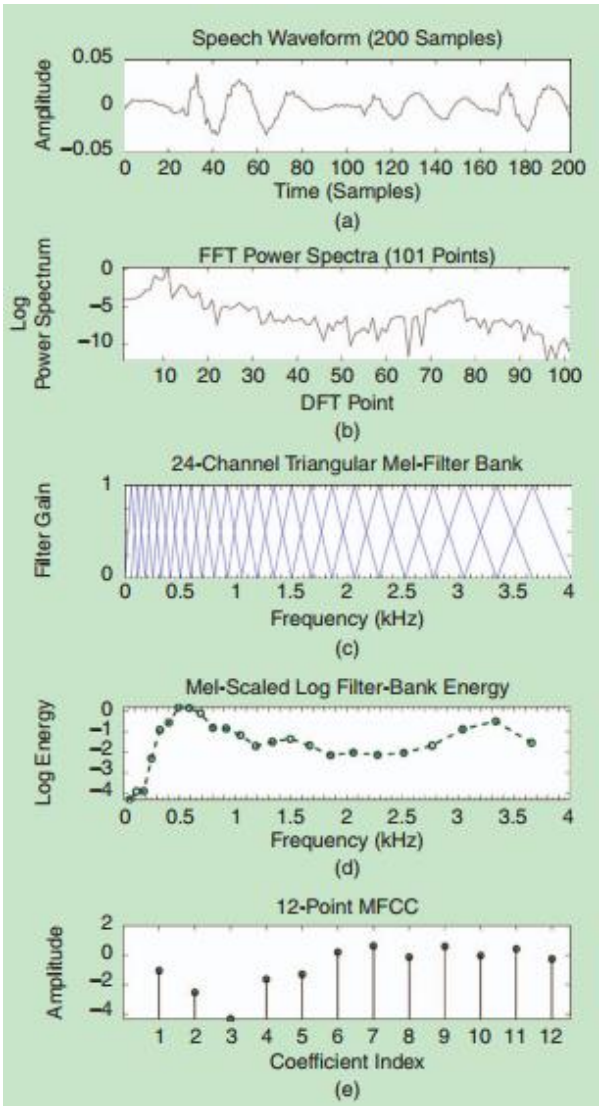

Figure 8: Steps in MFCCs feature extraction from a speech frame: (a) 200-sample frame representing 25 milliseconds of speech sampled at a rate of 8 kHz, (b) DFT power spectrum showing first 101 points, (c) 24-channel triangular Mel-filter bank, (d) log filter-bank energy outputs from Mel-filter, (e) 12

static MFCCs obtained by performing DCT on filter-bank energy coefficients and retaining the first 12 values (Hansen and Hasan 2015).

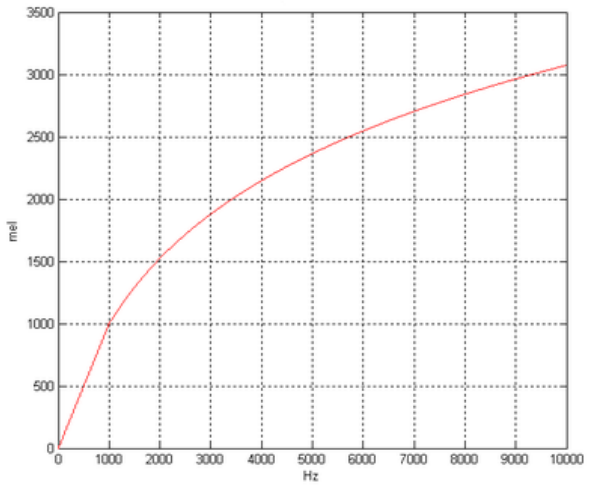

Figure 9: A plot of the mel transform. Horizontal axis: the frequency in Hz. Vertical axis: the corresponding mel values.

For examples, a doctor may need to be authenticated to access a reserved patient record from the desktop computer in the silence of his own office, or in a crowded loud department through a mobile handheld, or even through a simple voice call. In order to achieve robustness, i.e. low error rates, a huge amount of optimizations have been proposed by researchers all around the world, not only at the feature extraction level, but also throughout at architectural level and at each building block.

Anatomy and functioning of an Automatic Speaker Recognition system.

samples coming from many individuals, which represent the background population, bringing information about the variability of speech among different individuals (inter-speakers variability). In a next phase, the models of single authorized speakers can be trained, taking advantage of the background model computed in the first stage. Once the background and the authorized speaker models have been trained, the recognizer is ready to infer the identity of the incoming speakers (the "test" phase). In all phases, the voice samples must be captured and digitally converted, pre-processed (e.g. denoised), then entered in the feature extraction block, which typically produces a time series of features vectors. In each training phase, the features are processed to get a model of the speaker, which is stored in a database. In the test phase, the features of the test speaker are matched against the model(s) of the enrolled speaker(s) stored in the database for authentication (or identification).

So, after the feature extraction, a key role is played by the speaker modelling, i.e. the computation of a summary description of the feature distribution of a given speaker, and the related matching strategy. Many models have been proposed, such as the Vector Quantization technique, VQ (Burton 1987), where the feature vectors of the enrolled speakers are used to learn an optimal clustering (a codebook), specific to each speaker, of the feature

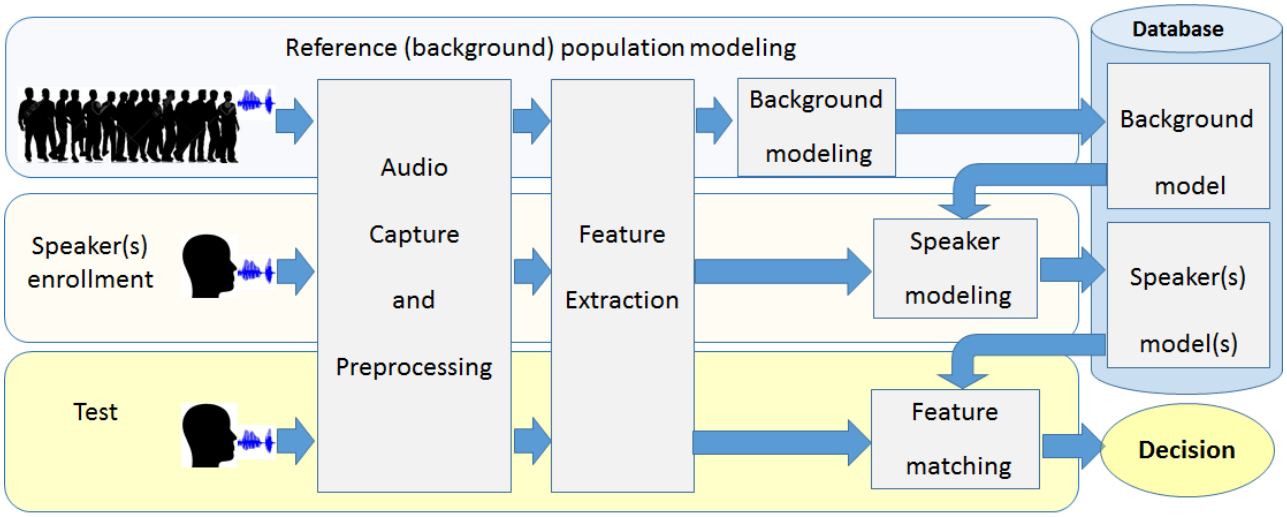

Figure 10: Schematics of a speaker recognition system.

Figure 10 shows a schematic of a typical speaker recognition system. In a first phase, the system needs to be trained with a number of voice

space. However, a first breakthrough came when the probabilistic approach was introduced, with the idea of modelling the variability

of each speaker with a probability density given by a mixture of multidimensional Gaussian components (GMM, Gaussian Mixture Model, Reynolds and Rose 1995), trained applying an Expectation Maximization (EM) algorithm.
In a basic identification task, the system returns the identity of the enrolled speaker whose GMM exhibits the highest statistical likelihood, given the feature vectors of the test speaker. In order to address the authentication task, a GMM is also trained with a large set of feature vectors coming from the utterances of many speakers, in order to get a so-called "Universal Background Model", UBM), for a reference population of speakers (Reynold et al. 2000). In this case, the feature vectors of the test speaker could be used to compute a likelihood ratio (LR) between the GMM of the claimed identity and the UBM, while the final response, Accept or Reject, could be provided comparing the LR to a threshold value.
A second breakthrough came with the introduction of the "supervectors", fixed-length vectors built by stacking the centroids of the GMMs components, that could be used in other processing steps, such as Support Vectors Machines, SVMs (Campbell et al. 2006). The latter is a machine learning classification technique that aims at optimally separating multidimensional data points (as the supervectors) from two classes (e.g. the enrolled speaker and the reference population) using a hyperplane, as depicted in Figure 11.
After the learning phase, if the supervector of the test speaker falls in the region associated to the claimed identity, the speaker is accepted.
Supervectors opened the doors to further improvements, because they made possible to apply Factor Analysis (FA), such as the Join Factor Analysis, JFA (Kenny 2005, Kenny et al. 2007) to separate (and discard) estimated contributes of noise and channel from that of the speaker, and, at the same time, to allow a some dimensionality reduction that make more practicable computational expensive tasks such as SVM training.

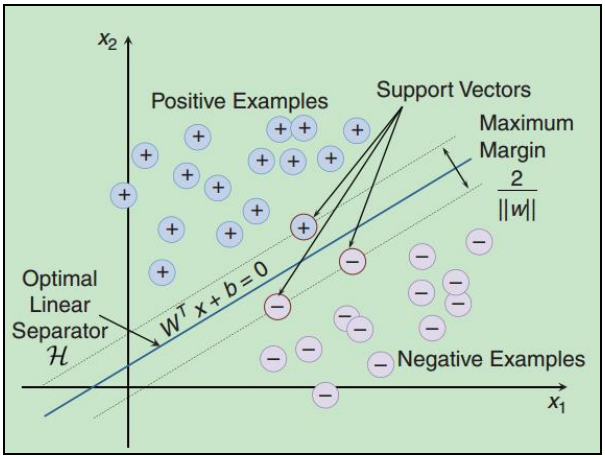

Figure 11: A conceptual illustration of a SVM classifier (Hansen and Hasan 2015). When applied to speaker identification, positive (+) points represent the vectors of utterances of the claimed identity, while negative (-) points represent utterances of the speakers belonging to the background population.

Moving from JFA, which was the yesterday state-of-art, some researchers introduced a new low-dimensional speaker- and channel-dependent space, defined using a simple factor analysis. The so called Total variability Space modelling (TVS, also known as "i-vectors" approach, cfr. Dehak et al. 2009 and 2011) provided a way to more effectively summarize the speakers utterances allowing at the same time a further dimensionality reduction and the use of noise and channel compensation methods that were not practical in large dimensional supervectors, such as Linear Discriminant Analysis, LDA (Vogt et al. 2008), Nuisance Attribute Projection, NAP (Campbell et al. 2006b), Within-Class Covariance Normalization, WCCN (Hatch et al. 2006). In the current state-of-start, the GMM-UBM/i-vector approach is combined with Probabilistic Linear Discriminant Analysis, PLDA (see Kenny 2011, Garcia-Romero and Espy-Wilson, 2011), a technique first used for session variability compensation for facial recognition, which works in principle as factor analysis in the i-vectors space (instead of the supervectors space). This allows an even further effective dimensionality reduction, which results in a more effective training phase and a simpler implementation. In the GMM-UBM/i-vectors/PLDA framework, the recognition is finally achieved comparing the LR of two alternative hypothesis, i.e. that the speaker-dependent component of the i-vectors of the test speaker and the claimed identity belong to the same identity, or not, with a threshold value.

The current state of art is rapidly moving toward Artificial Neural Networks (ANN), in replacement or in addition to some module in the traditional architecture. Deep Neural Networks (DNNs), large feed-forward ANNs with several (roughly 5-7) hidden layers and thousands of nodes, have recently been successfully used for acoustic modelling for the Automatic Speech Recognition task, achieving large improvements compared to standard GMM models. Researchers in the field of speaker recognition are proposing to use DNNs, and other kinds of ANNs such as the Convolutional Neural Networks (CNNs, see McLaren et al. 2015), as replacement of the UBM module (leading to a DNN/i-vector framework; see Lei et al. 2014), as de-noising module, as feature extractor module, leading to EER values up to 1-2% mainly in clean conditions, with an improvement over the traditional approach up to 30%.
Although in the field of voice biometrics there is still a lot of room for improvements, as for other biometrics technologies, the current state-of-art is ready to be successfully used in many sectors, including healthcare.

Voice biometrics for healthcare.
As already mentioned, biometric identifiers can be anything from fingerprints, palm veins, and iris scans to gait, facial, other than speaker recognition. Each one has its own benefits and disadvantages, but the latter has unique features that make it ideal for the healthcare sector, such as:

- Familiarity: using their own voice to interact with devices is becoming more and more familiar as voice-activated personal assistants and mobile apps spread.
- Little hardware required and cost effectiveness: any device equipped with a microphone and a network connection (voice or data) is ready. Voice biometrics user interfaces require no expensive scanning devices, only existing phone systems and everyday technology.
- No training is required to use a voice biometric system, because speaking is a natural activity.
- Mobility: voice biometrics works with any telephony technology, from anywhere in the world. Moreover, it is the only practical biometric option for authentication/identification over telephone channel.
- Accuracy: improvements in speech digital signal processing, such as noise cancelling and voice activity detection, as well as machine learning and deep learning are shrinking error rate under 1% in the most operative conditions.
- Effectiveness: as voice biometrics is actually a combination of both physical attributes (the shape of one's vocal tract) and behavioural attributes (speaking style), with a single interaction it provides authentication as well as "liveness testing." When used with speech recognition, voice biometrics can also be used to gather additional information such as a passwords, PIN, user-acknowledgement of prompted questions and so on, if required, with little or no extra cost.
- Integration with existing security systems: voice biometrics can be used with existing authentication methods to provide high-security multi-factor user-authentication.

Thanks to these concurrent properties, voice biometrics is an optimal solution, both as a stand-alone strategy and combined with traditional security approaches or other biometric identifiers for critical security needs, to improve the quality of healthcare services reducing costs at the same time.
Not only doctors but in general the healthcare staff as well as patients can get used easily with voice authentication and identification because recognition can happen in a seamless way, as if they speak at their own voice-activated personal app.
Electronic patient records can be safely retrieved by authorized personnel only, everywhere in the medical facility, by means of ordinary tablets and smartphones, other than desktop personal computers.
Simply using the intercom at the entrance to a department, a ward or any reserved area, could unlock access for authorized people, as the most natural gesture.
At-home healthcare services appear a perfect fit for voice biometrics and remote authentication,

that can perfectly work through a simple connection to landline or mobile phone networks.
In addition, the combination of voice biometrics and speech recognition unlocks many other opportunities, as voice assistants are entering the healthcare. For instance, patients can use such new devices as conversational interface, to get information and request something needed (see http://www.healthcareitnews.com/news/special-report-ai-voice-assistants-have-officially-arrived-healthcare). Within the same communication channel, doctors can both authenticate themselves and deploying accurate speech-to-text capabilities to transcribe notes and prescriptions, avoiding illegible handwriting and insufficient documentation of procedures (http://www.rightpatient.com/blog/voice-assistants-technology-healthcare/).

Conclusions

In the healthcare sector, biometrics are generally used in combination with traditional identification approaches (password, PIN, tokens) to authenticate/identify patients and staff, to improve security, cost containment and quality of service. Implementing biometric identification can address both, physical access control for computer/server rooms, hospital ward, and logical access control for login to the workstations and servers providing access to sensitive patient records.
Voice biometrics relies on the most natural form of human communication, requires only cheap hardware for user interface or no new hardware at all thanks to the large diffusion of personal mobile devices. Moreover, voice biometrics allows integration with existing traditional or biometric identification systems, improves mobility and allows at-home healthcare services. All this makes voice biometrics ideal for the healthcare sector in most use cases (except those cases where patients are temporary or permanently unable to speak, that can occur in emergency situations or in some hospital department).
Most of other biometric identifiers, also while successfully employed, require more or less expensive scanners, or cannot easily support remote (such as at-home) authentication), or are invasive with respect to the naturalness of the act of speaking.

Last but not least, voice biometrics technology has reached very high accuracy levels in most operating conditions, and can be successfully integrated in the state-of-art voice-activated personal assistant for many purposes, such as to boost patient engage

References

- Al-Kaltakchi, Musab T. S., Way L. Woo, Satnam S. Dlay and Jonathon. A. Chambers (2016), Study of fusion strategies and exploiting the combination of MFCC and PNCC features for robust biometric speaker identification, 2016 4th International Conference on Biometrics and Forensics (IWBF), Limassol, 2016, pp. 1-6.
- Burton, David K. (1987), Text-dependent speaker verification using vector quantization source coding, in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 35, no. 2, pp. 133-143, Feb 1987.
- Campbell , William M., Douglas E. Sturim, and Douglas. A. Reynolds (2006a) Support vector machines using GMM supervectors for speaker verification, IEEE Signal Process. Lett., vol. 13, no. 5, pp. 308–311, 2006.
- Campbell, William M., Douglas E. Sturim, Douglas. A. Reynolds, and Alex Solomonoff (2006b) SVM based speaker verification using a GMM supervector kernel and NAP variability compensation, in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Toulouse, France, 2006, vol. 1, pp. 97–100.
- Davis, Steven B. and Paul Mermelstein (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP28, pp. 357-366, Aug. 1980.
- Dehak, Najim, Réda Dehak, Patrick Kenny, Niko Brummer, Pierre Ouellet, and Pierre. Dumouchel (2009) Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification, in Proc. Interspeech, 2009, pp. 1559–1562.
- Dehak, Najim, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, (2011), Front-end factor analysis for speaker verification, IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 4, pp. 788–798, 2011.
- Demenko, Grazyna (2000) "Analysis of suprasegmental features for speaker verification", 8th Australian International Conference on Speech Science and Technology, 2000.
- Garcia-Romero, D. and C. Y. Espy-Wilson (2011) Analysis of i-vector length normalization in speaker recognition systems," in Proc. Interspeech, 2011, pp. 249–252.
- Hansen, John H. L., Taufiq Hasan (2015) Speaker recognition by machines and humans: A tutorial review, Signal Processing Magazine,

IEEE, vol. 32, no. 6, pp. 74–99, 2015.
- Hatch, Andrew O., Sachin Kajareka, and Andreas Stolcke (2006) Within-class covariance normalization for SVM-based speaker recognition, in Proc. Int. Conf. Spoken Lang. Process., Pittsburgh, PA, Sep. 2006.
- Islam, Md. Atiqul, Wissam A. Jassim, Ng Siew Cheok, Muhammad Shamsul Arefeen Zilany (2016), A Robust Speaker Identification System Using the Responses from a Model of the Auditory Periphery. PLoS ONE 11(7): e0158520.
- Kenny, Patrick (2005) Joint factor analysis of speaker and session variability: Theory and algorithms, Tech. Rep. CRIM-06/08-13, CRIM, Montreal, Quebec, Canada, 2005.
- Kenny, Patrick, G. Boulianne, Pierre Ouellet, and Pierre Dumouchel (2007) Joint factor analysis versus eigenchannels in speaker recognition, IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 4, pp. 1435–1447, 2007.
- Kenny, Patrick (2011) Bayesian speaker verification with heavy tailed priors, in Proc. Odyssey: The Speaker and Language Recognition Workshop, Brno, Czech Republic, 2010.
- Kim, Chanwoo and Richard M. Stern (2012), Power-Normalized Cepstral Coefficients (PNCC) for robust speech recognition, 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, 2012, pp. 4101-4104.
- Lei, Yun, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren (2014) A novel scheme for speaker recognition using a phonetically-aware deep neural network, in Proc. ICASSP, 2014.
- McLaren, Mitchell, Yun Lei, Nicolas Scheffer, and Luciana Ferrer (2014) Application of convolutional neural networks to speaker recognition in noisy conditions, in Proc. of **INTER-**SPEECH, 2014.
- Preite, Gianpasquale (2016), Politica e biometria: nuove prospettive filosofiche delle scienze sociali, Trento, p. 21, Trento, ISBN: 978-88-6458-164-4
- Purves, Dale. Neuroscience. Sunderland, Mass: Sinauer Associates, 2012.
- Ramos, Daniel Javier Gonzalez-Dominguez, Doroteo T. Toledano, and Joaquín González-Rodríguez (2015). Speaker Features, Encyclopedia of Biometrics, pp. 1455-1459.
- Reynolds, Douglas. A., Richard C. Rose (1995) Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE transactions on speech and audio processing, 3(1), 72-83.
- Reynolds, Douglas A., Thomas F. Quatieri, and Robert B. Dunn (2000) Speaker verification using adapted Gaussian mixture models, Digital Signal Processing, vol. 10, no. 1, pp. 19–41, 2000.
- Stevens, Stanley Smith, John Volkmann, Edwin B. Newman (1937). A scale for the measurement of the psychological magnitude pitch. Journal of the Acoustical Society of America. 8 (3): 185–190.
- van Leeuwen, David A., Niko Brümmer (2007) An introduction to application-independent evaluation of speaker recognition systems. In: Müller, C. (ed.) Speaker Classification 2007. LNCS (LNAI), vol. 4343, pp. 330–353. Springer, Heidelberg (2007).
- Vogt, Robbie, Sachin Kajarekar, and Sridha Sridharan (2008) Discriminat NAP for SVM speaker recognition, in Proc. IEEE Odyssey: Speaker Lang. Recogn. Workshop, Stellenbosch, South Africa, Jan. 2008.
- Yegnanarayana, B., S.R. Mahadeva Prasanna, Jinu Mariam Zachariah, and Cheedella S. Gupta (2005), Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system, in IEEE Trans. Speech Audio Process. , vol. 13(4), pp. 575-82, July 2005.
- Yegnanarayana, B., K. Sharat Reddy, and S.P. Kishore, "Source and system features for speaker recognition using AANN models," in proc. Int. Conf. Acoust., Speech, Signal Process. , Utah, USA, Apr. 2001.

Online resources

- Mayhew, Stephen (2015) History of Biometrics, http://www.biometricupdate.com/201501/history-of-biometrics. Last seen the 31st of January 2018.
- Gee, Jim and Mark Button (2015) The financial cost of healthcare fraud 2015: What data from around the world shows, PKF Lillteljohn LLP and University of Portsmouth, https://www.pianoo.nl/sites/default/files/documents/documents/thefinancialcostofhealthcarefraud-september2015.pdf
- Smart Card Alliance (2002) Smart Cards and Biometrics in Privacy-Sensitive Secure Personal Identification Systems, https://www.securetechalliance.org/resources/lib/Biometrics_and_Smart_Cards_Report.pdf
- http://www.healthcareitnews.com/news/special-report-ai-voice-assistants-have-officially-arrived-healthcare
- http://www.rightpatient.com/blog/voice-assistants-technology-healthcare