UNIL | Université de Lausanne
Ecole des sciences criminelles
bâtiment Batochime
CH-1015 Lausanne

Master of Science in Forensic Science

Physical Identification

Master Thesis

Detection of disguised speech in forensic science by humans and automatic systems

---

Michela Pettinato

Under the direction of Professor Christophe Champod
and the supervision of Doctor Sébastien Marcel

July 2020

Abstract

The step preceding the speaker identification process consists in the determination of the authenticity of a speech sample. The focus of this thesis is on the performance of humans in detecting altered samples from *replay*, speech synthesis (TTS) and voice conversion (VC) systems.

A listening test was constructed on the online survey platform LimeSurvey. The participants were asked to assess a series of recordings by first giving a binary evaluation ("authentic" or "altered") and then by specifying their level of confidence on a 5-point scale. Moreover, the logic behind the aural approach was studied by inspecting the criteria used by the respondents in the assessment of the recordings.

The same samples were also evaluated with an automatic LFCC-GMM-based system, trained on two different datasets, in order to make a comparison. The results show that the human's performance (EER=0.10) surpasses the machine's (EER=0.35 and EER=0.46) in the detection of the altered samples. However, sophisticated voice disguise systems have now reached levels of quality that can also easily fool most humans, which makes them a real threat to security and the identification process.

***

Keywords: Voice disguise, Biometrics, Speaker recognition, Authentication, Spoofing

Acknowledgements

First of all, I thank Prof. Christophe Champod and Dr. Sébastien Marcel for their support and their interest in my project. I would also like to thank Dr. Pavel Korshunov and Dr. Amir Mohammadi as well as the rest of the Biometrics Security and Privacy Group at IDIAP Research Institute in Martigny for their assistance and patience. A big thank you also goes to Prof. Romain Voisard for the help concerning the LimeSurvey tool and to my friend Annika Angeloni for the corrections.

Last but surely not least, I thank all the volunteers that allowed me to complete my project by participating in the survey.

Détection du déguisement vocal en sciences forensiques par des humains et des systèmes automatiques

Résumé

L'étape précédant le processus d'identification du locuteur consiste en la détermination de l'authenticité d'un enregistrement. L'objectif de cette thèse est d'étudier la performance des humains dans la détection des échantillons altérés provenant de systèmes de *replay*, de synthèse vocale (TTS) et de conversion vocale (VC).

Un test d'écoute a été conçu sur la plateforme de sondages en ligne LimeSurvey. Les participants ont été demandés d'évaluer une série d'enregistrements en donnant d'abord une évaluation binaire ("authentique" ou "altéré") puis en spécifiant leur niveau de confiance sur une échelle à 5 points. De plus, la logique de l'approche auditoire a été étudiée en examinant les critères utilisés par les répondants dans l'évaluation des enregistrements.

Les mêmes échantillons ont également été évalués avec un système automatique LFCC-GMM, entrainée avec deux sets de données différents, afin de réaliser une comparaison. Les résultats montrent que la performance des humains (EER=0.10) dépasse celles de la machine (EER=0.35 and EER=0.46) dans la détection des échantillons altérés. Cependant, les systèmes de déguisement vocal sophistiqués ont maintenant atteint des niveaux de qualité qui peuvent aussi facilement tromper la plupart des humains, induisant une menace réelle pour la sécurité et le processus d'identification.

***

Mots clés : Déguisement vocal, Biométrie, Reconnaissance du locuteur, Authentification, Spoofing

Remerciements

Tout d'abord, je remercie le Prof. Christophe Champod et le Dr. Sébastien Marcel pour leur soutien et leur intérêt pour mon projet. Je voudrais également remercier le Dr. Pavel Korshunov et le Dr. Amir Mohammadi ainsi que le reste du groupe Biometrics Security and Privacy de l'Institut de recherche IDIAP de Martigny pour leur aide et leur patience. Un grand merci également au Prof. Romain Voisard pour son aide concernant l'outil LimeSurvey et à mon amie Annika Angeloni pour ses corrections.

Enfin, je remercie tous les volontaires qui m'ont permis de mener à bien mon projet en participant au sondage.

Table of contents

1 Introduction

In recent years, a significant number of publications, both in the field of biometrics and forensic science, have been dedicated to the study of human and machine performance in speaker recognition.
Research indicates that nowadays the automatic system approach has, in many cases, reached or surpassed the auditory approach carried out by experts and laypeople (Hautamäki et al., 2010; Lindh et al., 2011). Some exceptions have been pointed out by Dessimoz (2004) and Wenndt et al. (2011): with more degraded samples (e.g. noisy recording settings) humans seem to perform better. These results justify a joint use of the two techniques of speaker recognition in criminal cases.

However, it is becoming increasingly important to focus on a very essential step that precedes speaker recognition: the determination of the sample's *authenticity*[1]. In forensic science, the question about the authenticity of any type of evidence is a concept of crucial importance. As specified by Maher (2018), the criminalist's evaluation of the sample stems directly from the circumstances creating it: if there has been some type of alteration, the investigator should be able to recognize it in order to assess the evidence correctly.
It is common knowledge that offenders often try to conceal their identity to avoid being recognized or impersonate an individual in an attempt to steal and misuse their identity. Both scenarios influence the forensic evaluation process and lower the confidence level that can be assigned to the collected evidence. In the context of speaker recognition, these types of situations, implying a manipulation of the speech evidence by the criminal, can be encountered in cases of telephonic fraud, blackmailing, kidnapping or bomb threats. The development of ways to detect those manipulations has become a central issue for experts in forensic science and biometrics, which are also putting the effort in correcting the overconfidence in speaker recognition, still considered to be an easy task by the general population (Clifford, 1980).

Because the terms used in biometric literature do sometimes diverge from those in forensic science, it is necessary to clarify them from the outset.

[1] Here, the term "authentic" does not hold the same meaning as in *biometric access control*, where the authentication consists in the process of verifying the claimed identity of a biometric system user. As defined in Pollitt et al. (2018) "in a forensic context, the claim to authenticate can include the integrity and provenance of a trace as well as contextual descriptions and temporal restrictions" (p. 6). The evidence can, therefore, be considered authentic if it has not undergone any kind of tampering or alteration.

Rodman (1998) defines the term "deliberate voice disguise" (p. 9) as being any conscious change in the voice of a speaker for concealment or impersonation reasons. In a biometric context, those actions are referred as “presentation attacks “, defined by ISO/IEC (2016) in the following manner: “The attacks to be considered in ISO/IEC 30107 are those that take place at the sensor during the presentation and collection of the biometric characteristics. » (p.1). In short, a presentation attack consists in the fooling of the sensor of a biometric system by falsifying the biometry presented to it. In literature, the term *spoofing* is often used to define a particular type of presentation attack in which an imposter circumvents the biometric system by using a counterfeit biometric of a third party in order to usurp their identity (Nixon et al., 2008).

Technically straightforward disguises, such as the use of *falsetto*, the obstruction of the nostrils, whispering or the use of a marked foreign accent, are considered to be among the most frequently encountered techniques when the goal is the concealment of identity (Masthoff, 2013; Perrot et al., 2012). Although relatively unsophisticated, these types of disguise still negatively impact the performance rates in speaker recognition tasks, both under the auditory and the automatic approach.
The next decade is likely to witness the appearance of much more refined disguise techniques, especially concerning the illegitimate impersonation of an individual. In the summer of 2019, two major newspapers, the Wall Street Journal and The Washington Post, wrote about a fraud case which was judged by experts to be "one of the world’s first publicly reported artificial-intelligence heist" (Harwell, 2019; p.1). The articles report on a fraudulent request for the transfer of a large sum of money from an England-based company to an account in Hungary. The crime was allegedly carried out with a software enabling the criminals to replicate the voice belonging to the general manager (Harwell, 2019; Stupp, 2019).
This incident shows that sophisticated impersonation techniques such as the ones using artificial intelligence (AI) are the new frontier of vocal disguise: they present a significant danger as they still seem to be very difficult to detect. Although the more refined voice disguise techniques seem to be less accessible to the general public, it must be noted that an increasing amount of readily available information and toolkits can be found on the Internet (ISO/IEC, 2016).

There are important differences in the estimates published by different law enforcement institutes concerning the number of evidentiary recordings presenting a deliberate modification of the voice. For example, the German Bundeskriminalamt (BKA) reports 15-20% of the assessed recordings as presenting some form of voice disguise (Künzel, 1994), while the Institute of Criminal Research of the French National Gendarmerie (IRCGN) estimates the number of cases as being 2.5% (Perrot et al.,

2012). Finally, Masthoff (2013) reported that in Germany, 52% of criminals tend to modify their voice if they suspect to be recorded. The variation between these estimates is an important indication of the difficulties that still surround the detection of these types of evidence manipulation.

While the automatic discrimination of authentic and deliberately disguised voice samples has been the subject of extensive research, to date only a limited number of studies have addressed the topic of auditory methods. However, this problem has recently sparked great interest among researchers (Farrús, 2018; Kamble et al., 2020; Wu et al., 2016), especially in the forensic science context, where speech evidence assessments are often carried out by humans. In 2016, INTERPOL published a survey conducted on the approach used by law enforcement agencies around the world when dealing with speaker identification. According to the paper, the auditory approach is used by 22% of the participating agencies on a worldwide scale. In Europe, 32% of the agencies base their evaluation solely on the human ear (Morrison et al., 2016).

This study calls into question the auditory detection of deliberate voice disguise in an *impersonation* setting. The aim is to study and compare the performance of humans and machines in order to better understand the level of confidence that can be attached to speech evidence under these approaches. Firstly, the voice disguise techniques will be introduced, followed by an examination of existing studies. Finally, the methodology used for the aural and the automatic approaches as well as the yielded results will be discussed.

2. Impersonation techniques

When the aim is to reproduce the characteristics of the speech uttered by a third party (impersonation), notably in the example of telephone fraud discussed above, several techniques presenting varying levels of sophistication and complexity are available.
In 1998 Rodman proposed a broad classification of the voluntary disguises which is still relevant in today's literature (Farrús, 2018; Perrot et al., 2012): he defined the techniques as being part of the "non-electronic deliberate disguises" or the "electronic deliberate disguises".

2.1 Non-electronic deliberate disguise

The only form of non-electronic impersonation is imitation. It consists in the process whereby a speaker modifies his voice in order to reproduce characteristics similar to those found in a third party's speech.

This can be achieved by adapting the pitch register, the dialect or even the voice quality (Perrot et al., 2005).

Several studies have been published on the ability of imitators to deceive automatic recognition systems as well as human listeners. Lau et al. (2004) show that the vulnerability of automatic systems increases if the natural voice of the speaker is similar to the target's, while Zetterholm (2004) determined that machines and humans are both likely to be deceived by a professional imitator.

Different results were obtained by Mariéthoz et al. (2006): the automatic system showed no vulnerability to professional, amateur or layperson imitators. Similar conclusions have also been expressed by Vestman et al. (2019): four laypeople were asked to reproduce the voices of different celebrities, but the disguise did not cause a significant deterioration in the speaker recognition task carried out by humans and machines. It is not yet possible to determine with complete clarity whether the ability to detect voice disguise by imitation is superior in humans or machines: this may be caused by the fact that the performed studies are based on very small quantities of data (Sahidullah et al., 2019).

A further discussion of non-electronic deliberate disguise falls outside the scope of this paper.

2.2 Electronic deliberate disguise

More recent is the appearance of techniques allowing to replicate the voice of an individual electronically, by using different degrees of technical skill.

Three types of electronic disguise are known in the forensic context: *replay*, speech synthesis, also known as text-to-speech (TTS), and voice conversion (VC) (Sahidullah et al., 2019).

Although more complex to create than non-electronic disguises, criminals can now take advantage of the help provided by several off-the-shelf open-source toolkits available on the Internet. The options include the Merlin speech synthesis toolkit (Watts et al., 2016), the Mary (Modular Architecture for Research on speech Synthesis) TTS toolkit (Schröder et al., 2011) or Lyrebird (Descript, 2017). This last software promises high-quality results based on a 1-minute-long voice recording of the targeted person.

2.2.1 Replay

The *replay* is the simplest type of electronic disguise. It consists in the reproduction of pre-recorded speech or speech fragments through a playback system (System 1 in Figure 1).

In essence, the voice is recorded by the criminal with an acquisition device, for example the microphone of a cellphone, and played to another system (System 2 in Figure 1) through a presentation device, possibly a cellphone loudspeaker (Evans et al., 2014; Sahidullah et al., 2019). It is also possible to inject the recorded speech directly into the system using a digital copy, without therefore needing a microphone

in System 2. The first method is called a *physical attack* (PA) while the second one is a *logical attack* (LA).

A limitation of this technique of vocal disguise consists in the potentially very limited recorded content the person owns of an uncooperative target. For example, in order to successfully commit a fraud of the type described above, there would be the need to obtain the recordings of some specific keywords. Moreover, the quality of the *spoof* highly depends on the quality of the used devices.

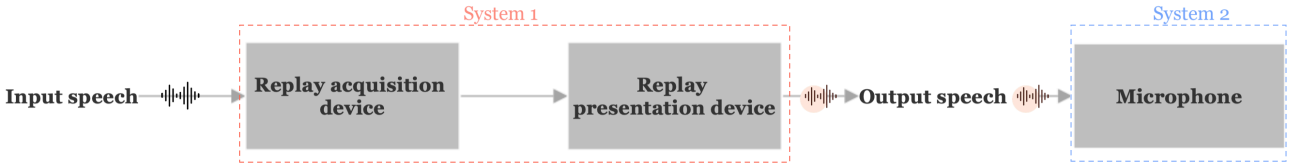


Figure 1: Block diagram of a replay disguise.

2.2.2 Speech synthesis (SS) or Text-to-speech (TTS)

Speech synthesis consists in the creation of artificial speech for any chosen written text. This process is possible by recovering the characteristics of a target's voice and then, as the term text-to-speech indicates, apply them to an arbitrary text determined by the offender (Kamble et al., 2020; Sahidullah et al., 2019). TTS technology is also used in many lawful applications, such as e-book readers, robotics or GPS navigation systems (Evans et al., 2014). Speech synthesis systems are typically composed of two main elements, also called *frontend* and *backend* (Figure 2).

Figure 2: Block diagram of a TTS system (adapted from Schroeter, 2008).

The first part provides a text normalization and linguistic analysis, where the introduced text is broken down into elements such as phonemes (consonants, semi-consonants and vowels). Then, the backend generates speech waveforms from the information provided by the frontend (Sahidullah et al., 2019; Schroeter J, 2008). The state-of-the-art TTS, in particular the use of deep-learning techniques such as in Wavenet (van der Oord et al., 2016), make it possible to create intelligible and natural-sounding vocal samples.

2.2.3 Voice conversion (VC)

The VC technique in forensic science consists in the conversion of the criminal's natural voice into the voice of another individual, without changing the content of the utterance. This process can be accomplished by using the target's voice as input, similarly to the text input in TTS (Evans et al., 2014).

As shown in Figure 3 and extensively explained by Wu et al. (2012), in a first stage the input speech signal is analyzed and several of its parameters are extracted. One of them is the fundamental frequency, also known as *pitch*. A transformation function previously trained on a large dataset of speech signals is then used in order to convert those parameters and, finally, a waveform for the converted speech is reconstructed by running said parameters through a *synthesis filter* generating the signal.

Figure 3: Block diagram of a CV system (adapted from Wu et al., 2012).

Numerous approaches and algorithms exist and, due to the rapid advances in deep-learning and the constant releases of new types of VC, it is difficult to define which of them is the most effective (Wang et al., 2019).

3 State-of-the-art: aural and automatic electronic disguise detection

3.1 Aural *spoof* detection

The study on electronic disguise detection by aural approach is still limited; however, there is interest in this particular topic and some papers have been published in recent years.

A study done by Wester et al. (2015) investigates for the first time the human ability to detect TTS and VC *spoofs*. The paper considers the real-world scenario in which the speech is transmitted through a wideband (16kHz) or a narrowband (8kHz) telephone line. Wester et al. (2015) used a sub-set of the SAS database, which includes five speech synthesis systems and eight voice conversion systems (Wu et al., 2015b). In order to replicate the telephone line, the original 16kHz data was downsampled and filtered.

In the conducted detection task, 60 English-speaking participants were given the following scenario: "*Imagine an impostor trying to gain access to a bank account by mimicking a person's voice using speech technology. (...) Your challenge (...) is to correctly tell whether or not the sample is of a human or of a machine*" (p.3). The listeners completed the test in a sound-isolated booth, using Beyerdynamic DT 770 PRO headphones. A total of 130 samples were randomly selected for each person: in order to minimize the bias in the listeners, 50% were authentic and 50% altered. The results of the study show that a higher detection error rate was recovered in the evaluation of 8kHz data (20%) than in the evaluation of 16kHz data (12%). Additionally, humans generally performed worse than tested automatic systems, except for one type of speech synthesis using *waveform concatenation*. This SS technique is a

widely used system that works by concatenating segments of authentic human speech recordings (Wester et al., 2015).

In a detection task presented in Wu et al. (2016), a total of 84 native English listeners were asked to discriminate human and artificial samples from a set of 130 utterances, of which 50% were authentic. The samples were taken from the SAS database. The same scenario and headphones as in Wester et al. (2015) were used and listeners were given ten examples which did not cover all possible alterations.

The results of this study are coherent with the ones obtained by Wester et al (2015). Wu et al. (2016) conclude by pointing out the interest in combining the decisions of both aural and automatic systems.

A study published in 2018 by Amino et al. used a corpus of ten native Japanese speakers with an age average of 22.1 years. A SONY ECM-23F5 microphone and a Marantz PMD671 PCM recorder were used in an anechoic room and the altered samples were created using three types of commercial speech synthesis applications based on *waveform concatenation*. A total of 20 native listeners took part in the experiment, where they had to choose if a sample was natural or synthetic. In essence, participants were given Sennheiser HD650 headphones and were allowed to listen to each sample only once before evaluating it. Half of the listeners, which were already educated on phonetics or linguistics, retook the test after they were given an oral explanation of the SS concatenation technique. The results show that, on average, natural speech was detected in 94.3% cases, while SS was detected in 65.3% cases. The speaker's identity and the knowledge of phonetics of the listener seem to have a significant impact on the evaluation: people with some knowledge in phonetics performed significantly better than the naïve group. However, the given explanation on SS concatenation did not show a significant effect.

Finally, a recent study on *spoofing* detection by humans by Todisco et al. (2019) completed the paper resulting from the ASVspoof2019 Challenge, developed for biometrics experts from academic and commercial backgrounds. ASVspoof is a world-wide bi-annual challenge that first took place in 2015 and counted 154 participants in its latest edition. This initiative aims to promote the development of generalizable and reliable automatic systems that can distinguish *spoofed* speech from authentic speech. The submitted systems are compared based on their performance in the assessment of samples from a database created specifically for the challenge (ASVSpoof2019 Consortium, 2019).

In the last edition, Todisco et al. (2019) also assessed the ability of humans in detecting the altered samples in the ASVspoof 2019 database and compared it to the results obtained with the machines. Seven TTS, three VC and three TTS-VC *spoofs* were included in the human evaluation sub-set, and they were balanced out by authentic samples (50%). In total, 1150 samples in English were submitted by crowdsourcing to 1145 subjects. In order to motivate the participants, they were told the following scenario: "*Imagine you are working for a bank call center. Your task is to correctly accept only inquiries from human customers and to properly determine those that may be due to artificial intelligence as*

'suspicious cases that may be malicious' (…)" (p.20). Participants were also informed that not all artificially produced sounds have a very evident robotic sound, depending on the technique used to create the sample. The subjects were then asked to listen to a recording as many times as they wanted and score each of them on a scale from 1 to 10, where *1=Absolutely machine-generated* and *10=Absolutely a human-produced utterance.* A total of 40200 scores was obtained in the survey, and the results show that the exactitude of people's evaluation depends mainly on the *spoofing* techniques used. In essence, state-of-the-art TTS using neural networks can fool most humans, while other samples were perceptually more easily identified as being artificially generated (Annex 1).

The publications discussing the vulnerability of human listeners to voice disguise clearly show that the performance of the aural approach decreases significantly when the used techniques allow to achieve a very natural sounding synthetic speech. Consequently, the level of confidence that can be attached to a speaker identification by aural approach seems to need a recalibration: a comparison between a reference speech sample and an altered evidence sample could potentially lead to a miscarriage of justice.
A disadvantage of the cited studies is that they do not attempt to understand the logic behind the human approach: the criteria used by participants in the evaluation of the individual recordings are not known. Moreover, despite the existing interest, no one to the best of my knowledge has studied *replay* samples in this setting.

3.2 Automatic *spoof* detection

3.2.1 Working principles

Until fairly recently, automatic detection of voice disguise was impossible since no such systems existed and there were still no databases allowing to handle this type of evidence. The human approach was, therefore, the only one available (Eriksson, 2010). However, especially in the field of biometrics, this topic has been pushed in the foreground in the last few years and the phenomenon spearheaded with the ASVspoof Challenges. The addition of automatic *spoof* detection systems to biometric systems is now seen as absolutely necessary to strengthen their security.

As briefly discussed above, *spoofing* can be of two types: physical (PA) or logical (LA). In physical spoofing attacks, the criminal presents altered samples to the biometric system's sensor while in logical attacks the sensor is bypassed and the *spoofed* sample is introduced directly into the system (Wang et al., 2019). A *replay* disguise where the recorded speech is played to the system is a type of physical attack, while TTS and VC are logical attacks (Figure 4).

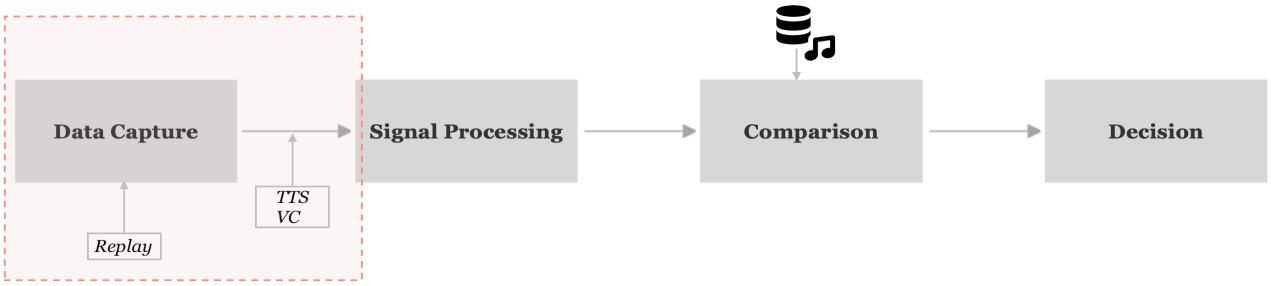


Figure 4: Block diagram of an automatic speaker recognition system with PA and LA points of attack (adapted from Muckenhirn 2019).

As explained in Anjos et al. (2017), the general way an automatic *spoofing* detection system works can be divided into three parts (Figure 5).

Figure 5: Block diagram of an automatic spoofing detection system (adapted from Anjos et al., 2017).

First, there is the preprocessing stage, where a voice activity detection system (VAD) is applied in order to exclude all of the non-speech segments from the sample before the feature extraction (Hansen et al., 2015; Mak et al., 2014). Non-speech segments primarily include silence or background noise.

Secondly, features are extracted from the speech segment. Different features are suggested in literature; in the AVSspoof challenges the most popular ones are the cepstral coefficient-based features such as Mel frequency cepstral coefficients (MFCC) and linear frequency cepstral coefficients (LFCC).

In the MFCC feature extraction the signal undergoes a segmentation into short overlapping frames of 20-45 ms, which are assumed to be stationary and independent. Then, a windowing of the signal segment *n* by a Hamming filter *w* of length *N* is performed in order to taper the signal at the edges and eliminate discontinuities between the multiple frames (Figure 6).

The Hamming window can be expressed as follows:

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)$$

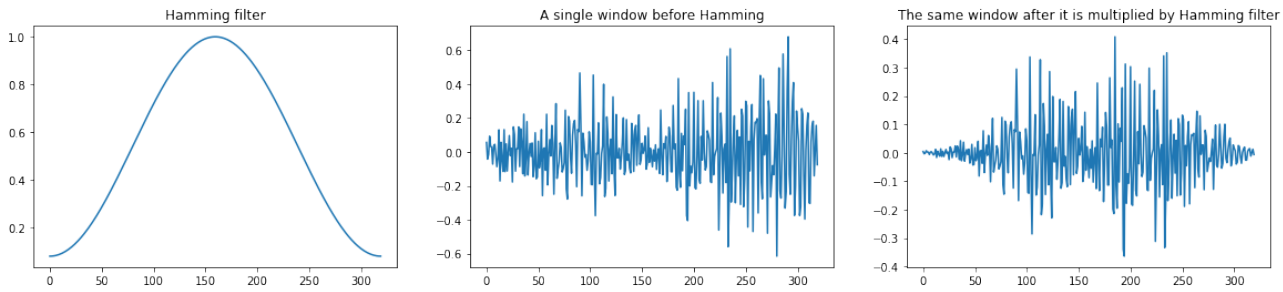


Figure 6: Hamming window (Korshunov, 2019).

The windowed speech is then transformed using Short-Time Fourier Transform (STFT) in order to convert the time-domain waveform into a frequency spectrum of the signal. Fourier Transform allows to detect which frequencies make up the speech signal, which can be very complex (Hansen et al., 2015). STFT, at the difference of the simple Fourier Transform, allows taking into account the inconsistency of wave signals in speech when creating the power spectrum.

Once the signal is transposed in the frequency domain, the information is compressed even more using the Mel-filter bank. This step consists in the application of triangular filters based on a Mel-scale, meaning a scale relating the *pitch*, the frequency as perceived by the non-linear human ear (f_{per}), to its real frequency (f_{Hz}) in the following way:

$$f_{per} = 2595\, log_{10}(1 + \frac{f_{Hz}}{700})$$

The manipulation of the signal allows to extract features which are close to those perceived by humans. This is achieved by having more discrimination at lower frequencies and less at higher frequencies (Hansen et al., 2015; Korshunov, 2019; Sithara et al., 2018).

The Mel-filter analysis allows to obtain the energy of the power spectrum in each filter-bank channel. The log of the found energies is run through a Discrete Cosine Transform (DCT) in order to decorrelate the obtained features and diminish their number even further. This whole process is repeated for each of the 25-45 ms long segments of speech and the extracted MFCC features can be stored in a matrix to create what is called a cepstrumgram, a very reduced form of the original signal.

LFCC features are similar to MFCC features, but with a difference in the placement of the filters: instead of being on a Mel-scale, they are on a linear scale (Sahidullah et al., 2015) (Figure 7).

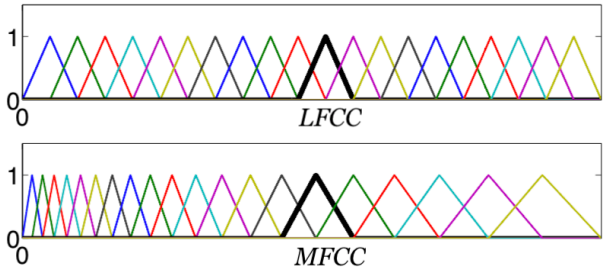


Figure 7: LFCC and MFCC filterbanks (Sahidullah et al., 2015).

It is important to note that the discussed features are not the only ones existing and being used. For example, Constant-Q Cepstral coefficient (CQCC) features have grown in popularity in recent years. CQCC features were introduced by Todisco et al. (2017) and are now outperforming most classical approaches. Instead of STFT, a Constant Q Transform (CQT) is applied to the waveform in order to get a greater frequency resolution for lower frequencies and a greater time resolution for higher frequencies. In essence, this allows an even higher resemblance to the human auditory system. As in MFCC, a DCT is applied in order to obtain the cepstral coefficients (Todisco et al., 2017; Wang et al., 2019).

Lastly, the *spoofing* detection system uses an algorithm called *classifier* on the extracted features in order to calculate and output scores, which will be used for the final decision: *spoof* or authentic. Scores can also be used in an evaluation framework in order to create plots and calculate error rates allowing to assess the performance of the used algorithm.

One popular classifier is the Gaussian mixture model (GMM), a mixture of *K* Gaussian probability density functions used to model multivariate data, often used in speaker recognition (Evans et al., 2014; Hansen et al., 2015; Muckenhirn, 2019). A vector x modelled by GMM is given by:

$$P(x|\theta) = \sum_{i=1}^{K} w_i N(x;\mu_i,\Sigma_i)$$

where x is the multivariate input $\in \mathbb{R}^d$, $\theta = \{w_i, \mu_i, \Sigma_i\}$, where w_i is the weight $\in \mathbb{R}$ with $\sum_{i=1}^{K} w_i = 1$, μ_i is the mean $\in \mathbb{R}^d$ of I, and Σ_i is the covariance matrix of i.

Each Gaussian is therefore characterized by a mean, a covariance matrix and a weight; the number of Gaussians needed depends on the nature of the available data.

With LFCC-GMM, CQCC-GMM is one of the state-of-the-art systems and both have recently been used as baselines in the ASVspoof2019 Challenge (Korshunov et al. 2016a; Wang et al., 2019).

In order to evaluate voice samples with an automatic system, the used algorithms need to be trained in advance using a data-driven approach. As explained in Mohammadi (2020), research datasets can be segmented in three non-overlapping parts, each with his own protocol: *training*, *development* and *evaluation.* Two possibilities exist when training and evaluating an automatic system: *intra-dataset* evaluation uses protocols from the same initial database, while in *cross-dataset* evaluation the training protocol does not come from the same dataset as the evaluated samples. In a real-world scenario like a forensic analysis, the setting is closer to a cross-dataset evaluation. In essence, it is not possible to know in advance which are the exact conditions of the recording the system will encounter during *spoofing* detection.

3.2.2 The performance of automatic anti-spoofing systems

In order to study the automatic anti-*spoofing* systems in the domain of speaker recognition, in 2015 the research community started releasing evaluation databases simulating different types of voice disguise. Those include, for example, the datasets used in the ASVspoof Challenges.

When evaluating and comparing automatic anti-*spoofing* systems, performance evaluation metrics are needed. In this paper, only the standalone assessment of the voice disguise detection system is discussed; it is also possible to evaluate both the speaker recognition algorithm and the anti-*spoofing* system using a metric like the tandem detection cost function (t-DCF) (Wang et al., 2019).

The most used metrics in biometrics are the false acceptance rate (FAR) and the false rejection rate (FRR), also called false positive rate (FPR) and false negative rate (FNR), respectively[2].

A false acceptance error occurs when *spoofed* speech is accepted as being authentic, and a false rejection error ensues when authentic speech is identified as being *spoofed* speech.

$$FPR(\tau) = \frac{\#\ spoofed\ samples\ with\ score > \tau}{\#\ total\ spoofed\ samples}$$

$$FNR(\tau) = \frac{\#\ spoofed\ samples\ with\ score \leq \tau}{\#\ total\ authentic\ samples}$$

It is possible to combine the two error rates to get the half total error rate (HTER), a more compact metric easily computable from FPR and FNR.

$$HTER = \frac{FPR(\tau) + FNR(\tau)}{2}$$

Usually, the error rates are calculated based on a decision threshold τ (Annex 2). As explained in Mohammadi (2020), some common criteria on which τ is chosen are: (1) a fixed value of FPR or FNR, (2) the equal error rate (EER) or (3) the minimum weighted error rate (min-WER).

When the FPR and the FNR equal, the performance is described by the EER: this metric, which defines a highly efficient system when very low in value, is used in the ASVspoof Challenges in order to compare the algorithms.

Meanwhile, Min-WER is a threshold τ that minimizes the expression:

$$cost * FPR + (1\text{-}cost) * FNR$$

where *cost* is the given weight. If cost has a value of 0.5, the criteria is called *minimum half total error rate* (min-HTER).

Finally, a powerful way to visualize the performance of one or more systems is the plotting of *receiver operating characteristic* (ROC) curves and histograms.

ROC plots are achieved by computing the false positive rate and true positive rate (1-FNR) for a range of possible thresholds τ. A perfect system would consist in 0% of FPR and 100% of TPR (Annex 3).

[2] In literature concerning anti-*spoofing* systems, the discussed error rates can also be called by their synonyms *attack presentation classification error rate* (APCER) and *bona fide presentation classification error rate* (BPCER): this new nomenclature was published in the ISO standards in 2017 (ISO/IEC). In this paper, FPR and FNR will be used, because this nomenclature is easier to transpose to the aural approach.

Recent studies on voice disguise show that it is still difficult to recreate realistic conditions and capture all techniques that can be found on the market. This means that the result in a real-world scenario could differ from the ones obtained in the labs (Wu et al., 2015a; Korshunov et al., 2016a; Kinnunen et al., 2017; Yamagishi et al., 2019).

Another important element that stands out from the published papers is the struggle to define which anti-*spoofing* algorithm gives the best results. Indeed, the performance of a system depends strongly on the nature of the disguise and can therefore fluctuate drastically.

Wester et al. (2015) compared the human performance, discussed in a previous chapter, to an MFCC-GMM based automatic system. The results for this system show an important variation in the error rates between the different disguises: the most successful automatic system preserves the naturalness of the voice by using the original waveforms to generate a *spoof,* while the least successful at fooling the machines introduces discontinuity into the speech by selecting and combining frames. The bandwidth also seems to significantly influence the disguise detection rates.

Wu et al. (2016) tested six automatic systems on five SS and VC *spoofs*: a total of six different types of features and two types of classifiers were compared. This study completes the human assessment by Wu et al. (2016) discussed in the previous chapter and uses the same dataset. For most *spoofed* samples, the system obtained FPRs below 1%, while humans had FPRs above 4%. Moreover, the results show that all of the tested systems are vulnerable to *spoofing* at different degrees and that the performance of the machine also depends on the nature of the used disguise and the data accessible to the criminal. Indeed, the greater the quantity and the better the quality of target speech data available, the more effective the disguise. According to the published paper, the global effectiveness of TTS and CV seems to be comparable and the difficulty consists in obtaining anti-*spoofing* systems able to detect them in a generalized manner.

In the most recent ASVspoof Challenge (2019), also discussed in the aural assessment chapter, two GGM-based systems were used as automatic system baselines: LFCC and CQCC (Wang et al., 2019). Especially for text-to-speech, the CQCC-features-based method showed lower absolute error rates when applied to the samples of the ASVspoof2019 dataset. However, the LFCC-based system seems to have a more stable performance across all disguise techniques (LA and PA spoofs).

In the ASVspoof2019 database, the PA *spoofs* are *simulated replay* disguises, obtained by adding different levels of reverberation and distortion to the signal, depending on the acoustic environment that was replicated. Moreover, the *replay* samples in this study were generated for different *talker-to-recognition system* (D_s) and *attacker-to-talker* (D_a) distances, and the use of multiple devices of different quality was taken into consideration (Annex 4).

In the assessment of said PA samples, the highest EERs are obtained for short D_a and D_s, high-quality replay devices and a low reverberation time: these particular conditions cause less background noise and distortions and are therefore more difficult to detect. The room size has little influence on the performance of the system.

Concerning the LA samples, the results obtained by the baseline systems show higher EER for VC systems than for TTS systems, especially when the voice conversion is based on Neural Networks, currently the most promising approach. The pooled EERs for all disguises in the evaluation set is 0.10 for the CQCC-GMM system and 0.08 for the LFC-GMM baseline.

Furthermore, EERs are reported to be higher for PA than LA, which shows that those particular automatic systems are fooled more easily by *replay* than by TTS or VC disguises.

In the LA scenario, 56% of the submitted projects outperformed the LFCC-baseline, while in the PA scenario the CQCC-baseline was bettered by 64% of the systems (Todisco et al.,2019). These results indicate that, even if the *spoofing* systems are more and more sophisticated, most of the samples presenting some type of disguise can be detected with state-of-the-art automatic systems. However, certain algorithms in the ASVspoof2019 database are still very difficult to identify automatically.

4 Experiment

4.1 Aim

The purpose of this experiment is to assess the human's ability to detect electronic deliberate voice disguise and compare it to an automatic system's performance. Additionally, the aim is to obtain information on which cues humans may be using when indicating that a speech sample is altered. This work adds to the existing studies by focusing on state-of-the-art voice synthesis and voice conversion techniques as well as *replay*.

4.2 Method

4.2.1 Aural approach

4.2.1.1 Databases

In order to choose a fitting corpus on which test people's skills, a careful selection of the appropriate databases is necessary. Firstly, the databases need to contain voice disguise techniques that could be used by criminals in present times and in the near future. In this study, the focus is mainly put on high-quality state-of-the-art techniques and simpler *replay* attacks. The former, even if currently seldom used, are the most refined and dangerous to an identification process. The latter are easily accessible to most

criminals and could therefore be more often encountered in police work, which makes them particularly interesting.

The databases used for the aural approach are described in detail below.

ASVspoof2019

This publicly available database was created for the ASVspoof2019 Challenge (Yamagishi et al., 2019; Wang et al., 2019). Although primarily designed to test biometric security systems, this database is also suitable for the forensic field. ASVSpoof2019 contains authentic samples as well as nineteen different TTS, VC and *replay* disguise techniques. The 107 recordings from native English speakers were produced in a hemi-anechoic chamber with a compact and omnidirectional microphone (DPA 4035). The frequency was reduced at 16kHz and the samples were stored in the *Free Lossless Audio Codec format* (FLAC).

One limitation for this specific database is in the *replay* settings: the *replay* samples in ASVspoof2019 are in fact authentic voice samples that were manipulated after the acquisition in order to simulate this particular type of disguise by adding echo and reverberation. This choice was based upon the need for creating carefully controlled setups where variables can be changed one by one (Wang et al., 2019). The resulting scenario is useful for studying biometric system's performances but far from the reality of forensic cases. For this reason, the set of *replay* samples was not taken from this database.

The original ASVSpoof2019 database, being too large for a study with human participants, was reduced to a sub-set by selecting two TTS algorithms and one VC algorithm, each presenting eight samples. Half of the recordings were female and half male. The decision on which of the several TTS and VC algorithms to choose was made based on two elements. Firstly, the possibility of coming across this type of disguise in a real criminal case and, secondly, the general danger it represents.

Concerning the first point, the most complex systems found in the ASVspoof2019 database are those combining TTS and CV (A13-A15). These *spoofs* were not considered, because technically more difficult to apprehend for the unskilled criminals, who also represent the majority of the population of interest. The second point, concerning the threat represented by the technique, was studied based on the ASVspoof2019 Challenge paper published by Wang et al. (2019). More specifically, the choice was made based on the results obtained by the two baseline systems using a Gaussian mixture model (GMM) back-end classifier paired either with *constant-Q cepstral coefficient* features (CQCC) or *linear frequency cepstral coefficient* features (LFCC) (Table 1).

Code (in Wang et al., 2019)	Type of disguise	EER (%)	
		CQCC-GMM	LFCC-GMM
A07	TTS	0.00	12.86
A08	TTS	0.04	0.37
A09	TTS	0.14	0.00
A10	TTS	15.16	18.97
A11	TTS	0.08	0.12
A12	TTS	4.75	4.92
A16	TTS	0.00	6.31
A17	VC	19.62	7.71
A18	VC	3.81	3.58
A19	VC	0.04	13.94

Table 1: Performance in terms of EER (%) for the ASVspoof2019 baseline systems (CQCC-GMM and LFCC-GMM). The disguises that were the best at fooling the automatic systems are highlighted in red (adapted from Wang et al., 2019).

The three chosen disguise systems are briefly described here:

A10. This is a state-of-the-art neural network (NN)-based text-to-speech system built by Jia et al. (2018). It allows to obtain a very natural voice presenting a high similarity to the target, from only a few seconds of audio.

A12. This is a NN TTS system based on the work of van der Oord et al. (2016), the neural waveform generator WaveNet. This model allows to produce waveforms of high quality that allow a natural-sounding speech similar to the target's voice.

A17. This is an NN-based VC system that was able to fool many systems in the ASVspoof Challenge 2018 (Huang et al., 2019; Wang et al., 2019).

The total of 24 altered LA samples was balanced out by adding 24 authentic recordings. This procedure is necessary in order to eliminate a perceptual bias introduced by an inhomogeneous distribution of samples (Wang et al., 2019).

ASVspoof2019_real_PA

In order to complete the evaluation sub-set chosen for this study, three types of *replay* settings were selected from the ASVspoof2019_real_PA database (Todisco et al., 2019).

This dataset of audio files in a FLAC format was publicly released in addition to the ASVspoof2019 database for the ASVspoof Challenge 2019. The small set of 2700 real *replay* audio files by 26 English speakers extends the existing simulated *replay* set in the first database.

The files were recorded in three different laboratories and the acquisition was made using a large number of different setups and environmental conditions (different device, room size, background noise, the distance between target speaker and recording criminal, …)

The original database being too large for a study with human participants, only three types of *replay* samples were chosen, each presenting eight samples, half of them being female.
The choice of the disguises was made based upon their feasibility and simplicity as well as the reality of criminal cases, prioritizing, for example, a medium-size office to a conference room and selecting moderately expensive, good quality acquisition and replay devices.

The three chosen replay disguises are the following:

PA1. Audio recorded in a medium-size office with an open window using a high-quality microphone (DPA 4035). The replay acquisition device used by the close-standing criminal was a MacBookAir built-in microphone and the replay presentation device a high-quality Bose soundlink III speaker.
PA2. Audio recorded in a medium-size office with an open window using a high-quality microphone (DPA 4035). The replay acquisition device used by the close-standing criminal was a high-quality BlueSnowball microphone and the replay presentation device a high-quality Bose soundlink III speaker.
PA3. Audio recorded in a medium-size office with an open window using a high-quality microphone (DPA 4035). The replay acquisition device used by the close-standing criminal was an iPhone 5S built-in microphone and the replay presentation device a high-quality Bose soundlink III speaker.

The difference between the three systems lays in the acquisition device: in order to better understand if the nature and quality of this element influences the assessment of the samples, the other variables were kept as a constant. Also, 24 authentic samples were added to the sub-set in order to create a balanced evaluation set: the authentic voice samples were all recorded with a DPA 4035 microphone in a medium-size office with a low air conditioning background noise.

Evaluation dataset for the study
The evaluation sub-set used in this study contains a total of 96 voice samples, 50% of which are altered (Annex 5). All files contain English utterances from native or non-native speakers and the mean length of the samples is 2.42 seconds (*SD=0.94 seconds*). Some examples of utterances are "*First meeting is next week*", "*I have had a lovely summer*" and "*My whole life has changed*". A well-designed short sentence of a few seconds can contain sufficient phonetic elements and be rendered naturally by the speaker (Wenndt et al., 2012). In order to be able to add the audio files to a survey interface, their format was changed from FLAC to the standard *Waveform Audio File format* (WAV). This transformation is lossless and does therefore not degrade the quality of the speech sample.

4.2.1.2 The survey

Participants

Of the initial cohort of 101 people that expressed their interest in participating in the survey, 82 completed and returned the questionnaire. The slight majority of respondents (59%) is female and the age range is between 18 and 63 (*Mean(age)=26*, *SD=8.08*).

The vast majority (94%) of those surveyed are studying or working at the Ecole des Sciences Criminelles in Lausanne: 29% are Bachelor students, 41% are doing their Master's and the remaining 30% comprises PhD's, Professors and Scientific Collaborators. The sample is therefore well balanced between the different actors of the forensic science faculty. The other 6% of volunteers work for the Biometrics Security and Privacy Group at the IDIAP Research Institute in Martigny, for whom the topic of biometric sample authentication also holds great interest. Since the aim of the study is mainly to identify how a police officer evaluates voice evidence, testing the general population was not of interest. Therefore, the sample will only include the two groups discussed in this paragraph.

The participants did not receive any type of compensation for the completion of the task.

Procedure

The survey was built on the online platform LimeSurvey (Schmitz, 2012), hosted on the university's server. This tool allows a secure use of the data and easy access for every respondent when using a token. A personal link to the survey, representing said token, was sent to all volunteers at the start of the study. This subjective experiment was uncontrolled, meaning that the conditions were not completely standardized and that the participants completed the task independently. The subjects could therefore use various types of listening devices. However, the listeners obtained the general indication to use headphones: this reduces the effects of the structural and acoustic properties of their environment, and therefore the variables in the subjects' experiences (ITU-R, 2019). Participants also chose the volume of the sound and listened to the samples as many times as they wished.

By using this survey's modality, several elements vary from case to case, influencing the perception of the respondents in different ways. This particular setting reflects the reality: conditions are not yet standardized for police officers listening to evidentiary recordings.

According to the ITU-R guidelines (1990), in order to reduce fatigue and distraction, a listening task should not last longer than 15 to 20 minutes. To limit the length of the survey, two different questionaries were created (A and B) by separating the evaluation set in two parts.

Each of the two documents contains a different set of 48 verbal expressions with the same amount of each type of disguise and the same ratio between female and male speech. Half of each set consists of

altered samples, while the other half contains authentic samples. Subjects that manifested their interest in participating in the survey were randomly assigned to one of the two question sheets. In the end, the same number of A and B surveys was completed. It is also important to consider that in a subjective assessment of sound quality at least 20 individual scores should be collected when the listeners are not experts in the field (ITU-R, 2019).

The participants were given the following scenario: *"You are a police officer listening to a voice recording obtained during surveillance. You have been advised that it is possible that someone is trying to frame the suspect and you must determine if you are confronted with an authentic sample of the suspect or if there has been any type of tampering. Try to detect which of the following voice samples have been manipulated to give the impression that the suspect is on the other side of the phone and which actually come from a conversation with the suspect".* Respondents were also asked to base their assessment solely on the sound properties of the recordings and ignore the contents of the utterances, which were inconsistent with the suggested scenario.
No further information was given regarding the nature of the possible alterations in order to avoid bias. In a real case, the police officer may not be acquainted with the existing state-of-the-art disguise techniques. However, four examples were presented to each volunteer at the beginning of the experiment. This was done in order to familiarize them with the samples: two of them were labelled as authentic and two of them as altered.
Moreover, the participants were asked if they were aware of being affected by some type of hearing impairment, which can be a limiting factor in the completion of subjective hearing tests (ITU-R, 2019), and they were questioned about their English skill level (A0-C1). The ITU-R guidelines for the subjective assessment of sound quality (2019) advise using native listeners when possible. Indeed, Hansen et al. (2015) show that in the speaker recognition task humans are more accurate when recognizing people speaking their own language; this may also be the case with disguise detection.

The experimental setup bears a close resemblance to the two-scale survey structure frequently used in witness memory studies, for example in Dodson et al. (2015) and Tekin et al. (2018). Initially, for each sample the participants were asked to answer a forced-choice question with a binary labelling. Forced-choice refers to a format in which respondents must provide one of two or more answers, in this case "authentic" or "altered", without having the possibility to abstain.
Immediately following this decision, they were instructed to indicate their confidence level on a 5-point scale, where *1=Not confident at all, 2=Slightly confident, 3=Quite confident, 4=Very confident* and *5=Practically certain* (Annex 6).

The choice for a 5-point scale lies in the fact that this type of rating system is the most used in human evaluation tasks and recommended in numerous studies. In essence, this scale gives more reliable results than the ones with finer granularity because it seems to be easier to understand and handle for subjects (Korshunov et al., 2015; Sinkowitz et al., 2013, van der Lee, 2019).

In order to limit bias in the data collection, several precautions were taken. First of all, as underlined by van der Lee et al. (2019), in the context of survey-based studies there may be a gradual or sudden change in the participants' work due to fatigue, an increase in confidence or other external factors.

The randomization of the question order for all different subjects, as well as an effort in limiting the duration of the experiment, can mitigate this effect.

Moreover, two *dummy* utterances were added at the start of each survey. These voice samples are of the same nature as the others in the questionnaire, however, they will not be included in the data analysis: participants require a short period to acclimatize with the task, which means that the first answers could be insufficiently reliable (De Simone et al., 2011).

At the end of the study, the respondents were asked on which general basis they decided to answer "authentic" or "altered", if they had detected different types of disguise and if they found the evaluation task to be difficult.

These follow-up questions allow to better understand the underlying logic of the aural approach in the detection of disguised voice samples.

Finally, in order to assess the sample evaluations made by the listeners, false positive rates and false negative rates were calculated, as well as the equal error rate. In the case of a false positive, it means that the listener incorrectly assesses an altered sample as being authentic. On the other hand, when the participant indicates that an authentic sample is altered, it's a case of false negative.

$$FNR = \frac{\#False\ negative\ scores\ ("Altered"\ when\ it\ is\ authentic)}{\#False\ negative\ scores + \#True\ positive\ scores} = 1\text{- true negative rate}$$

$$FPR = \frac{\#False\ positive\ scores\ ("Authentic"\ when\ it\ is\ altered)}{\#False\ positive\ scores + \#True\ negative\ scores} = 1\text{- true positive rate}$$

$$TER = \frac{\#False\ positive\ scores + \#False\ negative\ scores}{Total\ \#\ of\ scores}$$

Repeatability test

As explained in a 2014 guide about best practices in crowdsourcing (Hossfeld et al.), the consistency of the listener's rating behaviour needs to be tested in order to assure the reliability of the results.

The repetition of the test allows to determine if the scores obtained in the first survey are random guesses or if they are rooted in a more systematic evaluation process. However, Hossfelt et al. (2014) also warn

about the "familiarization and memory effects", which can influence the testing. In this study, the survey was repeated after at least one month had passed for each listener. Considering that studies on earwitness line-ups show that after three to four weeks the memory of a voice strongly degrades (Hollien, 2002) as well as the large number of samples to which listeners were exposed in the first questionary, it is safe to suppose that the "familiarization and memory effects" are no longer to be taken into consideration.
In order to decrease the variables between the two surveys, the participants were asked to use the same listening devices as in the first session.
In a span of two weeks, a total of 16 out of the original 82 participants (19.5%) retook the survey. Of those second-time participants, 43% are female and the age ranges from 20 to 63 years old (*Mean(age)=27.4*, *SD=10.11*). All of the listeners are part of the Ecole des Sciences Criminelles in Lausanne: 18% are studying for the Bachelor's degree, 44% are Master students and the remaining 38% are PhD, Professors or Scientific Collaborators. As for the first session, the participants were not remunerated for their work.

In order to process all of the aural approach data, the results obtained with LimeSurvey were exported as a *Comma-Separated Values* file (CSV) and then manipulated and analysed using the R programming language[3] (R Core Team, 2012). To execute the written code in batch mode and save it as an RMarkdown, the text editor RStudio[4] was used (RStudio Team, 2015) (Annex 17).

4.2.2 Automatic system approach

4.2.2.1 Databases

In this part of the research a combination of several databases is used in order to create the evaluation set and the training set. The first sample group, the evaluation corpus, contains all the voice samples already evaluated by humans (see 4.2.1.1) and is therefore built precisely as described in the aural approach. The second group is only necessary for the automatic systems and, in this case, a *cross-dataset* evaluation was performed.
The training process for the used model was carried out two times using different sets*:* the first one was built from a combination of the training sub-sets of the AVspoof (Ergünay et al., 2015) and SWAN (Ramachandra et al., 2019) databases, while the second one consists in the training sub-set of the ASVspoof2015 (Wu et al., 2015a) database. The results will allow to determine the impact of using different data for the training of the LFCC-GMM model.

[3] R Version R 3.6.3 GUI 1.70 El Capitan build (7735)
[4] RStudio Version 1.1.463

AVspoof

The publicly available database AVspoof, recorded at IDIAP Research Institute in Martigny over approximately two months (Ergünay et al., 2015), contains male and female authentic voice samples from 44 speakers as well as altered samples from the three categories found in ASVspoof2019: voice conversion, speech synthesis and *replay.* The data acquisition was made using different setups and environmental conditions: one good quality microphone (AT2020USB+) and two mobile phones (Samsung Galaxy S4 and iPhone 3GS) were used. All of the samples are in English and in a WAV format. VC was obtained by using a conversion function based on a Gaussian mixture model (GMM), TTS was based on statistical speech synthesis using the hidden Markov model (HMM), and the *replay* was created using different devices, more specifically phones and laptops with built-in or external loudspeakers (Ergünay et al., 2015). The training set of this database contains 51443 authentic and *spoofed* samples.

SWAN database

This multi-modal biometric database contains voice, periocular and face samples acquired using a smartphone application developed for iOS (Ramachandra et al., 2019). It contains the samples of 150 speakers collected in Norway, Switzerland, France and India; some of them are in English and some are in the native language of the participating laboratories. SWAN only contains *replay spoofs*, which were created by using two different high-quality loudspeakers and the microphone of an iPhone 6. The training dataset used in this study has a total of 4320 authentic and *spoofed* samples in a WAV format.

ASVSpoof 2015

The publicly available database ASVspoof 2015 was designed for the first ever ASVspoof Challenge and it contains authentic and altered English samples from a total of 106 speakers, of which 45 are male and 61 are female (Wu et al., 2015a). The training dataset contains a total of 16375 authentic and *spoofed* samples in a WAV format. All genuine samples were recorded without significant background noise. Concerning the *spoofs*, this database only contains LA attacks, more specifically ten different TTS and VC disguise techniques.

4.2.2.2 Automatic system used in the study

In order to evaluate the voice samples (*N=96*), the free signal processing and machine learning Python-based toolkit Bob 7 was used (Anjos et al., 2012; Anjos et al., 2017).

Bob provides a selection of readily available toolchains for signal processing and evaluation metrics as well as interfaces for accessing datasets: the environment was implemented using Conda, an open-source

package manager[5] (Anaconda Software Distribution, 2017). As explained in Mohammadi (2020), the toolkit Bob consists of over 100 packages and it is built upon the reproducible research philosophy by Anjos et al. (2017): it allows to create repeatable, shareable, extensible, and stable work. For this study, the following packages were used to run the experiment:

-`bob.bio.base`: base package which includes a generic script allowing to execute an experiment using one command line (Günther et al., 2016).

-`bob.bio.spear`: contains the tools needed for a biometric experiment on speech (Günther et al., 2016).

-`bob.db`: database interface for the used dataset (Anjos et al., 2017).

-`bob.measure`: contains functions for assessing the performance of the systems using metrics and graphical representations (Anjos et al., 2012).

-`bob.pad.base`: includes the basic definition of an anti-*spoofing* experiment and the necessary tools to run it (Anjos et al., 2017).

-`bob.pad.voice`: contains the tools needed to run *spoofing* detection for speech (Anjos et al., 2017).

Figure 8: Toolchain used in the experiment (adapted from Anjos et al., 2017).

The necessary toolchains (Figure 8) were already implemented in Bob, which allowed to perform the experiments efficiently and rapidly by using the fitting commands in the shell and the Python Application Interface (API) (Anjo et al. 2012). The LFCC-GMM system used and described here was tested by Korshunov et al. (2016a) in the paper *Cross-Database Evaluation of Audio-Based Spoofing Detection Systems*.

In the preprocessing phase, the system splits the sample into 20ms-long windows with 10ms overlap and uses a 4 Hz modulation in order to detect the spoken parts in the frames (Korshunov et al. 2016a). A typical speech signal possesses a frequency of around 4 kHz, hence, the VAD uses the modulation proprieties of speech to eliminate the parts that do not present this frequency (Maganti et al., 2006). The VAD logs can be inspected in order to distinguish how much of the sample is left to be analysed. In this study, two VAD parameters were tested: first, the default `no_filter` value, which keeps all frames, then the `trim_silence` value, which cuts the silent heads and tails of the signal (Günther et al., 2016). The extracted features are linear Mel-frequency cepstral coefficients and the employed algorithm is a 512 mixture GMM-based classifier with two models: one trained for authentic samples and one for altered samples. In the automatic system used in this project, each sample is compared to the two GMM

[5] Miniconda Linux 64-bit for Python 3.7

models and average log-likelihood scores are computed for $GMM_{authentic}$ and GMM_{spoof}. When the values are near 0 ((log(1)), the model and the utterance X are similar. On the contrary, the more the values go below 0, the less X and the GMM match (Korshunov et al. 2016b). Lastly, to obtain the final score $\lambda(X)$ for the sample, the ratio of the log-likelihood between the two GMM models is computed:

$$\lambda(X) = \log P(X|GMM_{authentic}) - \log P\left(X\middle|GMM_{spoof}\right)$$

where $X=\{x_1,\ldots,x_T\}$ is the feature matrix of the sample, T is the number of signal frames.
The higher this score, the more the machine will tend to identify X as being an authentic sample. However, the final classification depends on the decision threshold applied to the automatic system. In order to evaluate the systems, an EER is computed from the evaluation dataset containing the assessed samples. Finally, the routines in the package `bob.measure` were used in order to illustrate the performance of the systems through a ROC curve and histograms.

In order to process the data, the scores obtained from Bob were exported as a CSV file and then manipulated and analysed using the R programming language in RStudio (Annex 17).

5 Results and discussion

5.1 Aural approach

A total of 82 volunteers completed the two surveys A and B in about one month. An average of 20 minutes (*SD=11.20)* was needed by the subjects to answer all the questions.
The test resulted in a total of 3936 scores, 41 scores for each sample, which fulfils the conditions laid down in the ITU-R 2019 guidelines.
Only 4% of the participants indicated that they are aware of being affected by some type of hearing impairment; these people are included in the results because the effect on the data is minimal. Moreover, 91% of the 82 people participating consider themselves to be at least on a B1 English level (can make simple sentences and can understand the main points of a conversation). In essence, language is not considered a significant limitation in this study.
According to the results of a non-parametric Wilcoxon rank-sum test (Rosenberg et al., 2017) performed at a 5% significance level on the raw data, the difference between the scores assigned to the altered and the authentic utterances is significant (*p < 2.2e-16, for a confidence interval (CI)=95%*), meaning that the data produced by the humans should allow to classify the samples.

Indeed, the overall mean performance level, referring to the percentage of correct "authentic" and "altered" answers, is 75.5%, which is significantly above the 50% chance level for the two possible alternatives.

Of the total number of scores obtained in the forced question for authentic and altered samples, the mean false negative and false positive rates are respectively 21.8% and 27.3% (Figure 9). When combining the two errors, it's possible to obtain a total error rate of 24.5%, which, even though not extremely high, can influence the forensic evidence evaluation process negatively.

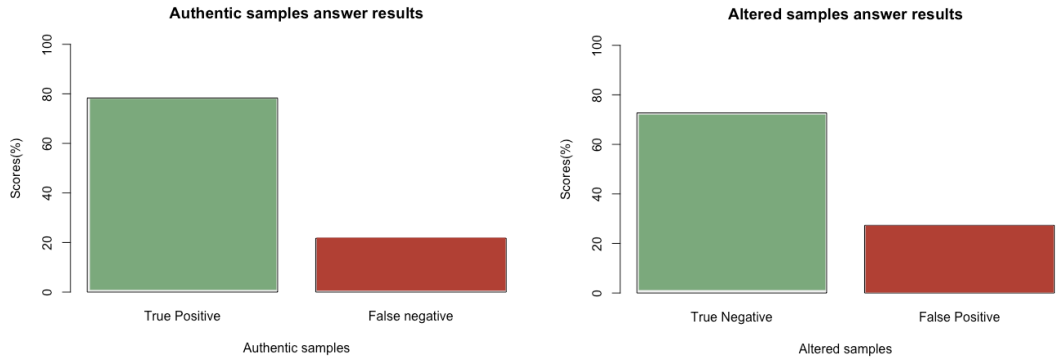


Figure 9: Barplots of the forced-choice questions' results for authentic and altered samples.

Authentic samples

The authentic sample data includes recordings from both ASVspoof2019 (logical access) and ASVspoof2019_real_PA (physical access). By simply observing the results it is possible to notice that both sets follow the same tendencies and, according to a Wilcoxon test, there is no significant difference between the two ($p= 0.12 >0.05$; $CI=95\%$).

First of all, the mean false negative rates are respectively 20.2% and 23.3%, meaning that the mean authentic sample detection rate (true positive rate) is at around 80%, which is largely above the chance threshold of 50%.

Secondly, in both sets there is a noticeable difference in the way participants assessed the individual samples: false negative rates go from very low (2% for LA and 5% for PA samples) up to around 45% (46% for LA and 44% for PA samples) (Annex 7). These results show that some of the utterances were clearly easier to classify than others. Interestingly, three of the samples showing the highest false negative rate (LA_E_3379393, PA_E_0048678 and PA_E_0053478) are recordings of female voices, while the samples with the lowest FNR are male voices. In order to try to better understand the difference in the assessments, some of the recordings are analysed in more detail. Table 2 contains four utterances with very high FNR and two with the lowest FNR.

Code	Length (sec)	Speaker gender	Transcript	FNR (%)	Analysis
LA_E_3379393	3	F	It's always nice to play on cinder court	46	Slight clicking noise at the end of the sentence, absent in most authentic samples.
PA_E_0048678	2	F	Mr. Smith was dismissive	44	"Flat" intonation.
LA_E_5849185	4	M	He has already suffered a good deal of unwanted attention	41	Very low pitch (F0=77.45 Hz)
PA_E_0053478	1	F	Then it will come	41	Fast speech, slightly raspy voice.
LA_E_7905661	2	M	The songs are just so good	2	Clear, slow speech. Emotional expression trough the intonation.
PA_E_0032774	3	M	However, we let them back into the game	5	Clear, slow speech.

Table 2: Authentic samples with the highest (dark grey) and lowest (light grey) FNR.

Only one male sample appears in the first four most incorrectly assessed authentic recordings: by listening to it, it is possible to notice that the pitch is extremely low. By performing a pitch analysis of the signal on the freely available voice processing toolkit Praat[6] (Boersma et al., 2020), it was possible to determine that the fundamental frequency (F_0) is 77.45 Hz. In Maher (2018) male talkers are reported to generally have a F_0 in the range 85-180 Hz. This "anomaly" in the sample may be the reason for the high FNR. The other samples at the top of the table also contain some elements that can explain their high error rate: electronic clicking noises in the background, unnatural intonation or unusual voice characteristics.

Authentic samples were best recognised when the speech was slow and clear and the speaker showed some emotion in his tone. Indeed, this last element seems to be difficult for electronic systems to reproduce.

Altered samples

Regarding the three different general types of disguise (TTS, VC and *replay)*, the first noticeable result is that globally all of them were detected by the majority of the listeners (Annex 7). Table 3 contains the mean detection rates for each specific *spoof*, with a standard deviation indicating the differences between the evaluated samples. The eight *replay* samples generated from one specific system (PA1, PA2 or PA3) seem to have been generally assessed more consistently than TTS or VC samples, which is made evident by the lower standard deviations. Moreover, the three different *replay* conditions yield similar results: this indicates that the exact nature of the replay acquisition device, as long as its quality is high, does not seem to influence the assessment in a significant manner (*all $p >> 0.05$, CI=95%).* In this study, a MacBook Air built-in microphone, a high-quality external microphone (BlueSnowball) and an iPhone built-in microphone were used. Other conditions, such as the room size, the background noise and the replay presentation device, were kept constant.

The disguise techniques A10 and A17 show the lowest and the highest detection rate respectively (27.4% and 99.1%). The TTS system A10 seems to have fooled most listeners into thinking that the samples are authentic: according to the results of a Wilcoxon test, the difference between the scores assigned to this

[6] Version 6.1.13

type of *spoof* and the authentic scores is only marginally significant (*p= 0.01 < 0.05; CI=95%*). The VC system A17 produces speech that was easily detected as being altered. These results confirm the previous research on the human assessment of the LA samples in the ASVspoof 2019 database (Wang et al., 2019).

Interestingly, one of the eight samples in the A17 set was sometimes incorrectly evaluated (LA_E_2085042): three participants out of the 41 (7%) mistakenly identified it as being authentic. The particular sample is a male, robotic sounding, utterance of the short phrase "*It's so awful*". The reasons behind this result are not wholly understood.

Disguise	**Type**	**Mean Detection rate (%)**
A10	TTS	27.4 (SD=13.5)
A12	TTS	80.2 (SD=10.7)
A17	VC	99.1 (SD=2.6)
PA1	Replay	77.5 (SD=5.3)
PA2	Replay	77.8 (SD=6.7)
PA3	Replay	73.5 (SD=6.7)

Table 3: Mean detection rates in the altered samples.

Confidence levels

The aim of this thesis is not only to assess the *spoofing* detection rates in humans, but also to study the level of confidence attached to the listener's decision making.

Firstly, it is of interest to investigate if the subjects exploited the totality of the confidence level scale or if they mostly favoured the extreme scores, as it was the case in Van Dijk's study (2013).

The results (Annex 8) show that "3- Quite confident" was the most used answer (30.4%), closely followed by the score "4-Very confident" (27.1%). The scores "5- Practically certain" and "2-Slightly confident" were used around 19.7% and 19.5% of the times, while the score "1- Not confident at all", was only used in 3.3% of the answers. This analysis demonstrates that the participants were generally reasonably confident in their answers and that they did not rely solely on the more neural scores, even if they are the most used.

The results also reveal that generally the participants were more confident when they correctly evaluated the authenticity of the speech: as seen in Figure 10, high scores (4 and 5) are more prominent in accurate answers, while low scores (1 and 2) are mostly encountered when the answer is incorrect.

However, especially in the case of false positives, the number of "very confident" and "practically certain" proves to be non-negligible (*N=132* and *N=59*, respectively). Also, the subjects sometimes indicated a low confidence level of 1 or 2 even when their given answer was correct: in 22 cases listeners indicated that they are "not confident at all" when they fittingly identified an authentic utterance and the same happened in 45 cases of true negative evaluation.

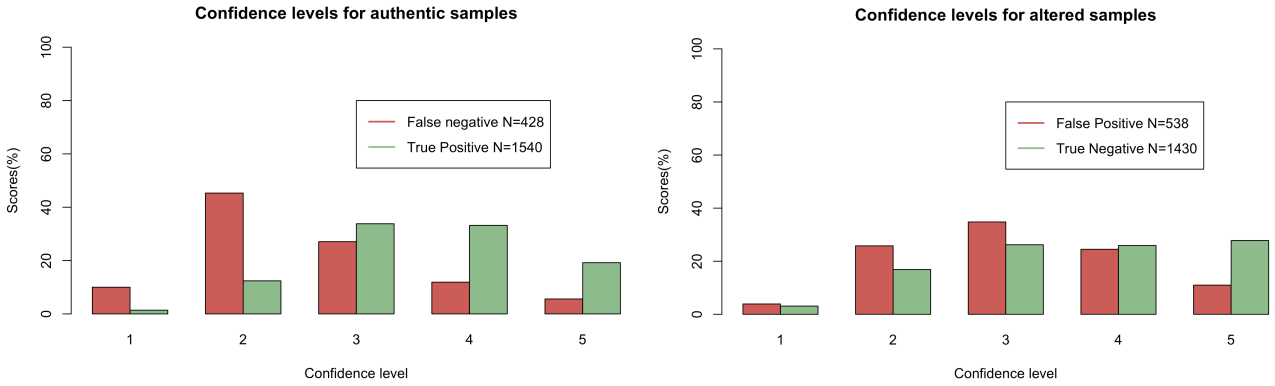


Figure 10: Confidence levels in cases of correct (green) and incorrect (red) evaluation.

In order to further analyse the survey's results, the scores of the forced answers and the confidence levels were combined. The approach is inspired by the one used in Van Dijk et al. (2013) *A human benchmark for automatic speaker recognition,* where two 5-points confidence level scales, one for the answer "same speaker" and one for "different speaker", were represented as scores from -4.5 ("certain: different") to 4.5 ("certain: same"). In this paper, the two 5-points confidence scales for "authentic" and "altered" were aggregated in a scale going from -4.5 (practically certain: altered) to 4.5 (practically certain: authentic). This way of representing the data allows to obtain a new scale symmetric around the value zero and containing all the possible answers to the survey.

The complete aggregated statistics are visible in Table 4.

	« Altered »					« Authentic »				
Confidence level	**-4.5**	**-3.5**	**-2.5**	**-1.5**	**-0.5**	**0.5**	**1.5**	**2.5**	**3.5**	**4.5**
Authentic	24	51	116	194	43	22	191	520	512	295
Altered	397	371	375	242	45	21	139	187	132	59

Table 4: Aggregated statistics, the incorrect answers are shown in red.

In this table, the most interesting values are 24 and 59: they represent the total number of incorrect answers for which the listeners were very confident in their decision. Those cases would very likely have ended in a false exclusion and a false identification, respectively.

It is of interest to investigate what all averaged scores indicate for the different samples and types of disguise. First of all, it is possible to notice that the average score for a specific sample is mostly an accurate assessment: all authentic samples have scores above zero (Figure 11), which indicates that globally the samples were correctly recognized as not having been manipulated. However, the participant's answers for most samples show a high variation (Figure 12), possibly caused by either different listening conditions or diverse skill levels.

Some of the authentic samples have a mean score that is noticeably lower or higher than the average score (*Mean=1.95, SD=0.66).* It is pertinent to listen to said recordings and try to detect the characteristics that cause them to be more difficult or easier, respectively, to assess. The details of the ten audio files with the highest and lowest scores are given in Annex 9. The authentic samples were best

recognised when the speech was slow and clear with no background noise and the speaker showed some emotion in his tone. Clicking noises, low pitch and raspy voices caused the smallest scores.

Concerning the altered samples, five out of the six types of disguise (A12, A17, PA1, PA2 and PA3) follow the same tendency as the authentic samples: globally, the recordings have average scores that are below zero (*Mean=-1.52, SD=1.73*), which means that generally they were evaluated correctly. However, as shown in Figure 11, the utterances created with the TTS system A10, which is part of the ASVspoof_2019 database and produces very good quality synthetic utterances, have high average scores (>0). The results show clearly that this type of *spoof* fooled most listeners, which identified the samples as being authentic with a high confidence level. The details of the ten audio files with the highest and lowest scores are given in Annex 10. The clear and natural speech created by A10 produced the highest scores, while the robotic voice from the A17 disguise scored poorly.

The participant's answers show the most variability in the *replay* samples scores (PA1, PA2 and PA3), while for the A17 samples the results are relatively stable between the volunteers (Figure 12).The A17 VC *spoof* produces a relatively unnatural sounding speech, quite easy to recognize as being altered, which explains the coherence in the individual answers.

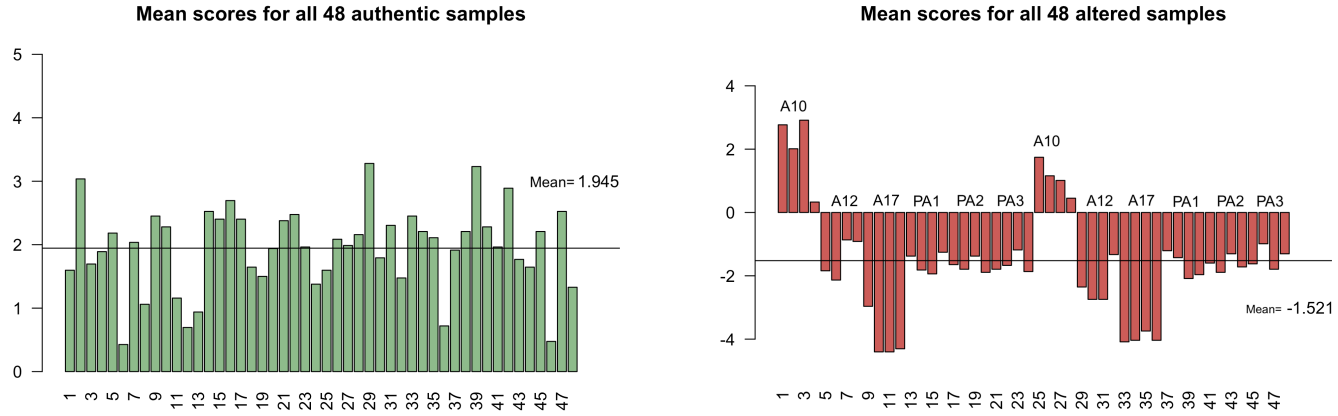


Figure 11: Barplots with the average scores for each authentic (green) and altered (red) sample. The type of disguise is specified on top of the bars.

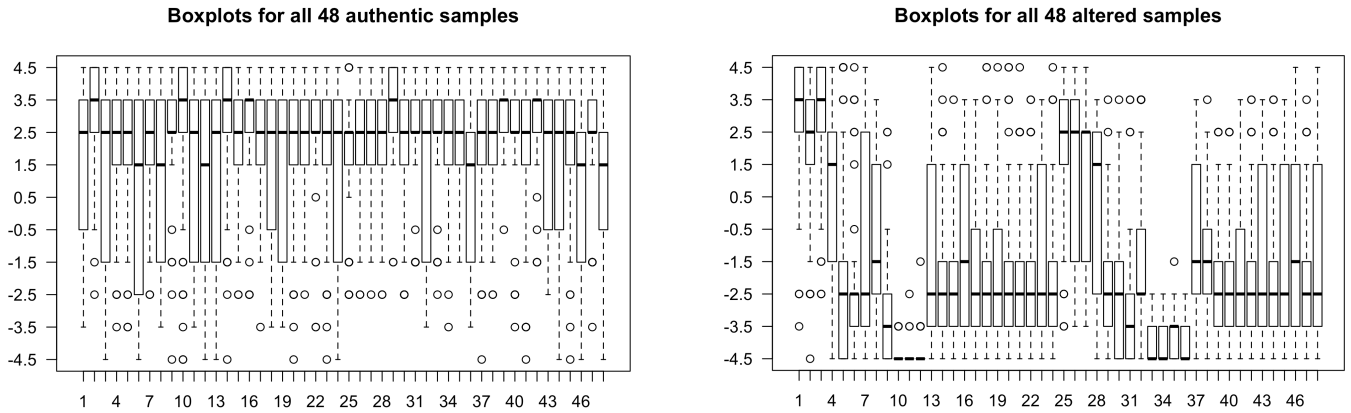


Figure 12: Boxplots for each authentic and altered sample.

With the package `bob.measure`, it was possible to plot the performance of the humans on a ROC curve (Figures 13, 14 and 15) and calculate the EER for the average sample scores. In this case, the low value of the metric (*EER=0.10)* indicates a good general aural assessment. The equal error rate for LA samples (TTS and VC) is EER=0.16, while for the *replay* samples it is EER=0, because all of them have been, on average, correctly identified.

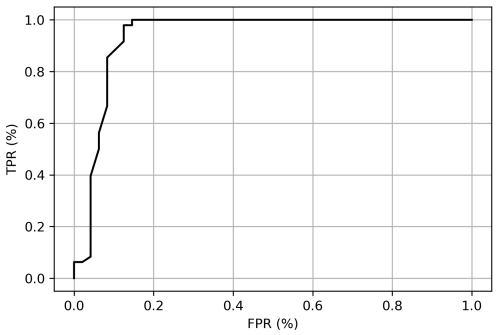
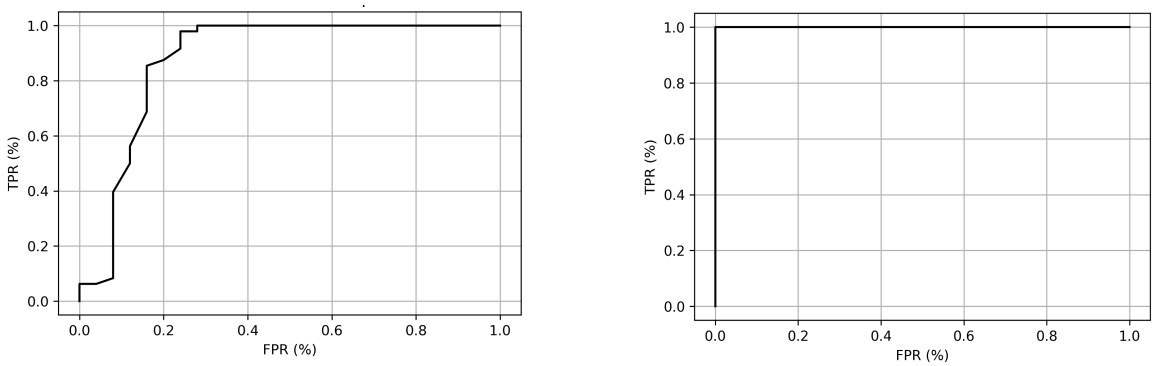


Figure 13: General ROC for the aural method (PA+LA).



Figures 14 and 15: ROCs for the aural method: TTS + VC samples (left) and replay samples (right).

Follow-up questions

Four follow-up questions were added to the survey to understand how the participants reached their conclusions regarding the authenticity of the samples.

In the first question, people were asked to describe on which basis they decided to label a sample as being "altered" (*On what basis did you choose samples which, in your opinion, are not authentic? Give one or two examples (e.g. electrical noise, distortions, pitch, discontinuity in speech, intonation,...*). From the given answers, it appears that a vast majority was motivated by the presence of all or some of the following elements when answering "altered": distortions, electric disturbances, robotic sounding voices, discontinuity or hesitation in the speech and persistent background noises.

Unnatural intonation, sometimes described as "flat" or "monotone" and occasionally as "illogical", was also pointed out by many of the participants. Another parameter that was often brought up is the presence of an "echo" effect in some of the samples (*replay*) which makes the voice appear as distant and more subdued. Some of the participants expressed their uncertainty in evaluating those particular samples,

explaining that they felt like the authenticity of those recordings cannot be ruled out if, for example, the speaker is located in a car or another small space. This element explains the differences found in the individual scores assigned to the PA samples (Figure 12).

The technical term *replay* was found in three answers, which indicates that at least three participants are aware of the existence of this particular type of disguise. In order to create a visual representation of the answers given, all of the responses were collected in a *text file* (TXT) and manipulated in order to obtain a WordCloud, picturing the most frequent words in a bigger size (Figure 16). The elements emerging from the answers should be recognized as being good indicators for *spoof* and used as such in casework.

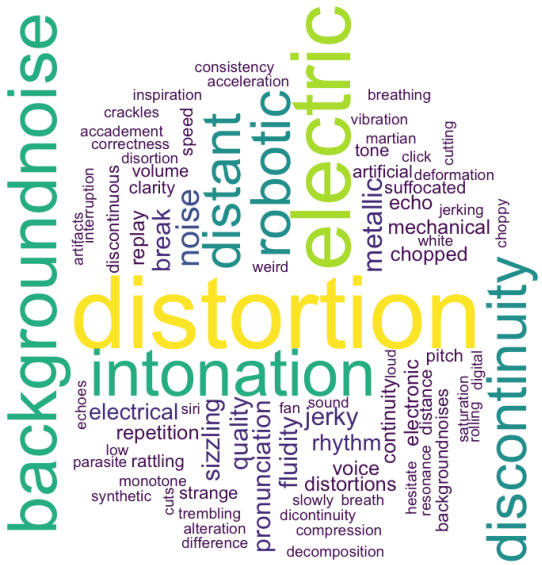


Figure 16: WordCloud representation of the answers.

In the second question, the participants were asked if they were able to detect different types of alterations, meaning different *spoofing* techniques (*Did you detect different types of alterations in the not authentic voice samples? If yes, describe how*). While a minority (17%) answered "no" or "not really", most of the responses are detailed and allow to determine how aware naïve subjects are to different types of deliberate electronic disguise.

First of all, it appears that people generally separated two types of samples: "robotic" or "distorted" speech, often described as "Siri voice" or "fake voice", and natural-sounding samples that however exhibit sharp background noises or echoes. The latter were sometimes explained as being "like recordings played from far away", and a few people mentioned the presence of a clicking sound at the beginning and the end of the sentences. Some listeners further discriminated between speech samples that they perceived as being completely synthetic and others who sound reasonably natural and seem to have undergone merely some minor manipulation. One good example is given by this particular answer: "sometimes the voice does not resemble a real voice (…), as if it was completely synthesised. Other times, it seems a real person's voice which was slightly modified". Another person writes that "there are

levels of alteration". However, nobody advanced the possibility of different algorithms being used and allowing therefore to obtain different qualities of *spoofed* speech.

In a limited number of answers (about 10%), more technical terms and concepts were brought up, for example, *replay* and "automatic reading of a text". Those answers reflect an underlying grasp of the general notion of deliberate electronic disguise and its techniques. Finally, some listeners indicated that some of the samples seem to have simply been reconstructed from pre-existing recordings of speech segments, which is not really the case in this experiment.

The third and fourth questions investigate if the respondents perceived the experiment as globally difficult and if female voices were more challenging to evaluate compared to male voices or *vice-versa (Did you find this exercise difficult? Did you feel that you had more difficulty with male/female voices or both equally?).* The results show that 76% found the task challenging: this demonstrates that even if the recognition of voices is an effort that humans perform daily, there is a level of uncertainty and doubt attached to the assessment of recorded speech samples. Most of the respondents (61%) indicated that female and male voices were both equally tricky to evaluate. The data does not show a correlation between the gender of the listener and the given answer (Figure 17).

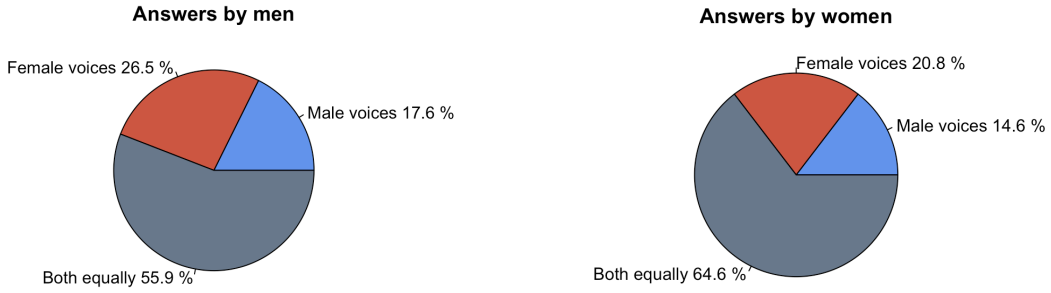


Figure 17: Answers to the follow-up question "Which samples were more difficult to evaluate?".

In order to determine if the survey's results reflect the answers given by men and women to the question on which samples were more difficult to evaluate, the rates of incorrect « authentic » and « altered » answers were calculated for both female and male samples. More specifically, the false positive rate (FPR), the false negative rate (FNR) and the total error are determined (Table 5).

Sample / Respondent		FPR (%)	FNR (%)	Total error (%)
Female	Woman	25.5	26.6	26
	Man	31.1	22.5	27
Male	Woman	28.5	18.6	24
	Man	24.5	18.6	22

Table 5: Comparison between female and male samples for female and male respondents.

From the table it is possible to see that the difference in the total error is not significant: both genders seem to assess female and male voices in a similar manner. These results corroborate the participant's responses to the follow-up question. The genders of the listener and the speaker do not seem to influence the assessment.

Repeatability test

A total of 16 out of the 82 original volunteers (20%) completed the surveys for a second time. The differences between the two session's CSV are analysed in RStudio using the *comparedf* function from the CRAN package *arsenal* (Heinzen et al., 2020).

An average of 14 minutes (*SD=3.62)* was needed by the subjects to answer all the questions and the repeatability test resulted in a total of 768 new scores.

The yielded results show that in total 45.4% of the answers were changed by the listeners from one session to the other, however only in 26.3% of cases the general decision ("authentic" or "altered") was changed: in most instances solely the indicated confidence level varies. From the results it is possible to determine that the *replay* disguise has the highest number of changes in the assessment. Indeed, in 32% of the cases the listeners answered the binary query differently from the first session, making the results less reliable.

The absence of complete repeatability in the answers is coherent with the fact that most of the listeners indicate that they found the task challenging (63%) and this result needs to be considered when speaking about the strength of this type of evidence in court. Nevertheless, it is also possible to notice that the difference in the general FNR (20% and 23.7%) and the FPR (27% and 26.8%) between the two sessions is not significant, meaning that the overall performance did not change substantially.

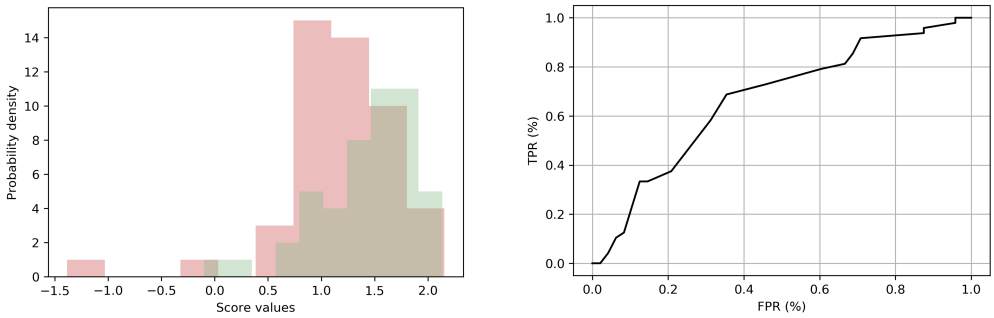
5.2 Automatic system approach

LFCC-GMM system trained on the AVspoof+SWAN dataset (Training 1)

From the score distribution histogram plotted in the Python API (Annex 11), it can be observed that the system does not allow a separation between altered and authentic recordings. The equal error rate (EER) was calculated on the evaluation set using `bob.measure` and has a value of 0.38.

It is also possible to notice that one of the samples behaves as an outlier: it is a female 4-second long utterance from the disguise technique A12 (*score: 9.82*). The file in question (LA_E_3005039), contains the sentence "*The military campaign is a campaign against Osama bin Laden*". By listening to the audio file, it is possible to hear silence at the end of the recording: the hypothesis of this characteristic being the cause of the observed behaviour was tested by adding the `trim_silence` filter to the VAD system.

The results yielded after the implementation of the filter, illustrated in Figures 18 and 19, were created with the `bob.measure` package.



Figures 18 and 19: Histogram and ROC for the automatic system (AVspoof+SWAN).

The influence of the non-speech part proved to be important, therefore all further analysis were made on the scores obtained selecting the speech segments only. The results show a higher separation of authentic and altered samples and the system's performance increases slightly (*EER=0.35).* The average score for the authentic samples is 1.43 (*SD=0.44*) and 1.17 for the altered samples (*SD=0.57*); according to a non-parametric Wilcoxon rank-sum test performed at a 5% significance level on the raw scores, the difference between the data is significant (*p=0.01<0.05, CI=95%*). However, the FNR and FPR at the EER threshold are elevated, causing the approach to be highly unreliable. This can represent a serious problem in forensic science as well as in biometrics.

Authentic samples

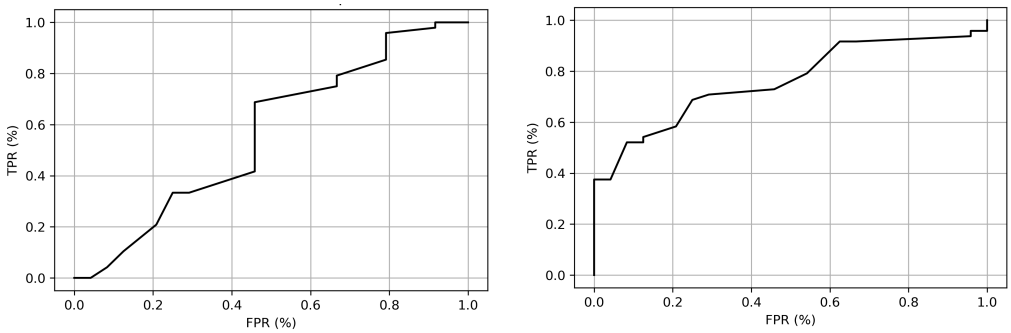
All of the authentic samples have a score above zero, except for PA_E_0114962, a 2-second-long female utterance ("*He will go a long way*"), which presents a score of -0.10. PA_E_0114962 and the 2-second male utterance LA_E_8739004 ("*Your father was a good man*") (*score=0.33)* were the only two samples incorrectly identified as altered. By simply listening to these two files, it is not clear why the system misclassified them. More detail on the five authentic samples with the highest and the lowest scores are found in the Annex 12.

Altered samples

The mean score for all *replay* samples is 1.13 *(SD=0.25)*, while the mean score for the TTS and VC samples is higher, at a value of 1.21 *(SD=0.77)*. The automatic assessment of the *replay* samples in the different settings seems to be more consistent, which can be seen from the smaller standard-deviation. From the obtained results, it is understood that the difference in the *replay* acquisition device used influences the way the system assesses the samples: PA1 has a mean score of 1.41 (*SD=0.12),* PA2 a mean score of 1.10 (*SD=0.19)* and PA3 a mean score of 0.90 (*SD=0.11)* ; the differences are significant *(all p<0.05, CI=95%)*.

The system's performance for PA samples (*EER=0.29)* is superior to the performance for TTS and VC samples (*EER=0.46*), meaning that *replay* was slightly less difficult for the machine to detect (Figures 20 and 21). However, if the considered threshold is the pooled EER=0.35, most of the samples are incorrectly assessed as being authentic (Table 6). Indeed, only two utterances from the disguise A17 (LA_E_4361221 and LA_E_3659898) were correctly evaluated by the system at scores of respectively -1.38 and -0.19.

The first one is a male 4-second utterance ("*It looks as though he will be void (?)*") and the second one is a 3-second male utterance ("*He admitted he was attracted to women*"). Both have a very strong distortion causing the voice to sound unnatural.



Figures 20 and 21: ROCs for the automatic system (AVspoof+SWAN:) TTS + VC samples (left) and replay samples (right).

Disguise	Type	Detection rate (%) for τ_{EER}
A10	TTS	0.0
A12	TTS	0.0
A17	VC	25.0
PA1	Replay	0.0
PA2	Replay	0.0
PA3	Replay	0.0

Table 6: Detection rates (TNR) for the different spoofs

Surprisingly, the A17 sample LA_E_2085042 ("*It's so awful")* was given a very high score (2.16), which differentiates it from the other utterances in this type of disguise. Indeed, most of the other samples in A17 were given low scores, which means that the system is more likely to correctly classify them as being altered. One plausible explanation for this is that LA_E_208542 is preceded by a lengthy silence: after the addition of the `trim.silence` filter, not much was left to be analyzed, which increases the chances of error.

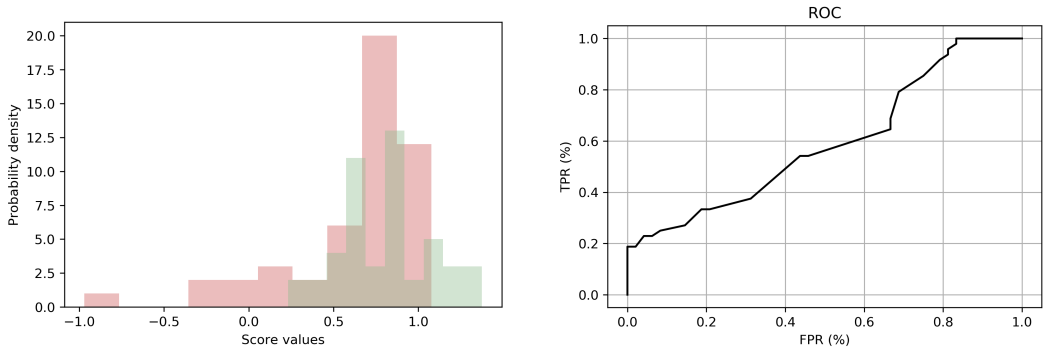
More difficult was the automatic detection of A12 and A10 disguises: this can be seen from the high scores attributed to most of said TTS samples. More detail on the five samples with the highest and the lowest scores are found in the Annex 13.

In conclusion, the LFCC-GMM system trained on the AVspoof and the SWAN datasets does not allow to recognize the altered samples well: if trusted, it can cause false identifications which can lead to a miscarriage of justice.

LFCC-GMM system trained on the ASVspoof2015 dataset (Training 2)

The first results obtained by using the ASVspoof_2015 database in the training phase, are illustrated in Annex 14. From the images plotted in the Python API, it can be observed that the system does not allow a separation between altered and authentic samples (*EER=0.50*).

As in the first training session with AVspoof and SWAN, the behaviour of the sample LA_E_3005039 (*score: -2.12*) can be explained by the presence of a long silence at the end. By removing the non-speech segments of the audio files, the results are the following (Figures 22 and 23):



Figures 22 and 23: Histogram and ROC for the automatic system (ASVspoof2015).

The results show a higher separation of authentic and altered samples: the system's performance increases slightly (*EER=0.46).* However, the error rates are still unreasonably high, making the automatic system highly unreliable. The average score for the authentic samples is 0.80 (*SD=0.26*) and 0.64 for the altered samples (*SD=0.40*); the difference in the data is not significant (*p=0.12 > 0.05, CI=95%)*.

Authentic samples

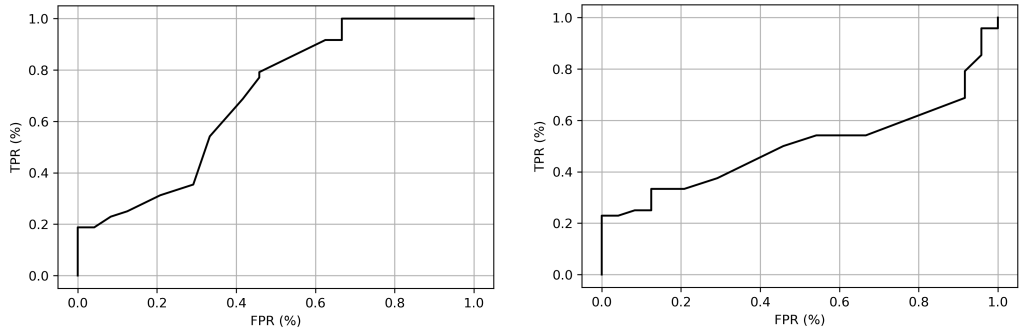
All authentic samples have a score above zero. By using the pooled EER=0.46 threshold, four samples are incorrectly classified as being altered. The utterances are all female, however, it is not possible to determine why the system has made errors with these particular audio files by simply listening to them. More detail on the five samples with the highest and the lowest scores are found in the Annex 15.

Altered samples

The mean score for all *replay* samples is 0.79 *(SD=0.13)*, while the mean score for the TTS and VC samples is lower, at a value of 0.47 *(SD=0.50)*. The assessment for the *replay* samples was more

consistent but also less accurate. One logical explanation for this is that there were no *replay* samples in the database used for the training of the GMM model. According to a statistical analysis using the Wilcoxon test, the *replay* acquisition device does not significantly influence the results (*all p>0.05, CI=95%*): PA1 has a mean score of 0.76 (*SD=0.17),* PA2 a mean score of 0.84 (*SD=0.10*) and PA3 a mean score of 0.77 (*SD=0.11)*. The system's performance for said samples (*EER=0.47)* is inferior to the performance for TTS and VC samples (*EER=0.39*), meaning that *replay* was slightly more difficult for the machine to detect (Figures 24 and 25).

All eight utterances from the disguise A17 were correctly classified by the system (Table 7). One possible explanation for this result would be that the same disguise technique was already employed in the ASVspoof Challenge in 2015 and was therefore present in the used training set. However, this is not the case, and therefore the high detection rates are possibly due to the fact that this particular type of algorithm produces unnatural sounding speech that can easily be recognized.



Figures 24 and 25: ROCs for the automatic system (ASVspoof 2015): TTS + VC samples (left) and replay samples (right).

Disguise	**Type**	**Detection rate (%) for τ_{EER}**
A10	TTS	0.0
A12	TTS	0.0
A17	VC	100.0
PA1	Replay	12.5
PA2	Replay	0.0
PA3	Replay	0.0

Table 7: Detection rates (TNR) for the different spoofs.

The assessment of one male 1-second *replay* sample from PA1 (PA_E_0086372) also resulted in a true negative ("*It was clear")* (*score= 0.38*). This particular recording contains a voice with a heavy south-Asian accent, which may have influenced the evaluation process. Indeed, the database used for training contains utterances mostly produced by native English speakers. More detail on the five samples with the highest and the lowest scores are found in the Annex 16.

Concerning the LFCC-GMM system used, it is possible to notice that the resulting performance is degraded compared to the one obtained for the ASVspoof2019 Challenge baseline LFCC-GMM (Wang et al., 2019). There are two reasons for this: the first is that the models used by Wang et al. (2019) are trained on a large number of databases, which are not necessarily the same as used here. This element is decisive in the sample assessment by automatic systems, as seen by the differences in the two experiments conducted here. Therefore, while it was very useful to get a comparison between machines and humans, the system presented in this paper does in no way pretend to be the best anti-*spoofing* system existing today. The second reason, which is the most obvious, is that not all samples of each *spoof* and not every *spoofing* technique from the challenge were used in this paper. Indeed, only a fraction of the possible disguises was selected. On the other hand, the ASVspoof2019 Challenge paper reports on the evaluation of all techniques (A7-A19), including the ones that produce low-quality speech and are therefore very easy to detect.

5.3 Aural vs automatic system approach

By comparing the results from the aural and the automatic approaches, it is evident that the volunteers participating to the survey, especially when considering the average scores that were obtained, performed significantly better (*EER=0.10*) than the tested LFCC-GMM systems (*EER=0.35* and *EER=0.46*). Humans correctly classified all *spoofs* except for one TTS (A10), while the automatic system was often incapable of separating authentic and altered speech.

By looking at the scores assigned to the different samples, it is possible to note that with both approaches A17 was generally the easiest disguise to detect. Its robotic features were recognized by humans and machines. On the contrary, the A10 *spoofing* system was difficult to identify, causing it to be it dangerous to the forensic identification process. However, while A10 was by far the technique that fooled people the most due to the high quality of the synthesized voice, the results are not as straightforward in the automatic systems. Indeed, sometimes A12 samples ranked higher than A10 in the machine approach, even if the utterance is less natural-sounding to the human ear. These observations are in line with the results previously published in Wang et al. (2019).

The PA samples (*replay)* were generally better detected than the LA samples (TTS+VC), exception made for the LFCC-GMM system trained on the ASVspoof2015 database, which did not contain PA samples. As in the earlier studies, the performance of both machine and humans greatly depends on the type of disguise and its capability to produce natural-sounding speech without important distortions or background noises.

In conclusion, this experiment shows that, with the selected state-of-the-art disguises (A10, A12, A17 and the three *replay* settings PA1-PA3), humans have more success in recognising the authenticity of an

utterance. Despite this result, this paper also offers compelling evidence on the limits of the speaker identification by auditory approach. Indeed, the individual participants were sometimes mistaken in their evaluation, hence, criminalists should always ask the opinion of a second expert and, if possible, combine the results with automatic systems' scores in order to increase the confidence level in the assessment.

6 Conclusion

The aim of this thesis is to study how skilled humans are at correctly recognizing disguised voice samples and compare the results with the performance of a *spoofing*-detection system. The purpose is to better understand the evidentiary value of speech recordings assessed under an aural or automatic approach.
The obtained results, showing a higher performance for humans than for machines, are in line with the findings in the recent speaker recognition studies. Indeed, while automatic systems perform generally well, humans are better at assessing utterances in difficult or degraded conditions.
When evaluating the recordings, the participants based their decisions on several criteria: altered samples were mostly associated with distortions, electric disturbances, robotic sounding voices, discontinuity and persistent background noises. These elements could be used as indicators for potential *spoofs* in future casework.
The auditory approach is promising when dealing with the new generations of electronic deliberate voice disguise, however, as already detected by Wang et al. (2019), some high-quality *spoofs* that can fool most humans already exist and constitute a great danger to the identification process.

Despite the general agreement with other recent studies, the findings need to be interpreted with care since the experiment may have some limitations that greatly depend on the databases that were used. Firstly, all of the authentic samples are of perfect quality, which does not replicate the real-world scenario accurately. An authentic recording would likely present distortions and background noise in a real case of interception. In this experiment, listeners could recognize most genuine samples simply by their exceptional clarity and lack of disturbances, which possibly influences their perceived vulnerability to *spoofs*.
Secondly, the used sentences were very short and the text was not freely spoken, meaning that the subjects were uttering a series of crafted scripts in a mostly neutral ton. Said condition also causes the experiment to only restrictively mimic what could be encountered in real casework.

Interestingly, the survey's follow-up questions allowed to expose a big uncertainty that some of the listeners experienced during the test: when confronted with *replay* samples, certain participants expressed their need to have access to more contextual information in order to assess them correctly. While the samples in question were mostly perceived as being different from those used as the "authentic" example in the beginning of the survey, a few of the listeners indicated that they would have considered the recordings as being authentic if there had been mention of them coming from the inside of a car or a small space. Indeed, those particular conditions would explain the "echo" and the "distant voice" effect that is typically found in *replay* samples. These types of comments underline the importance of contextual information in the process of evidence evaluation, as is the case for many types of evidence. Furthermore, this problematic helps to better understand the lack of experiments concerning the human evaluation of *replay*: the detection of this type of disguise is strongly influenced by the context and possibly the type of device used to listen to the samples. This last element needs further study, as already suggested by Wang et al. (2019).

The results obtained in the repeatability test show that over a quarter of the authentic-altered assessments (26.3%) change in the second session, meaning that the answers given are not always repeatable. This observation once again challenges the idea that great confidence that can be attributed to the conclusions given by humans in the process of speech evidence evaluation.

In a future study, it would be interesting to test the listeners in a controlled and standardized experiment, where each person has access to the same listening device and where their work can be supervised. Such circumstances would allow to limit the influence of the material used by the participants and consequently to compare the performance of the individual respondents.
Moreover, in order to better replicate the real-world scenario, background noises and distortions could be added to the authentic samples.

7 References

Amino K, Makinae H and Kamada T (2018) Auditory discrimination of natural speech and synthetic speech used as voice disguise. *Acoustical Science and Technology* 39(1): 48-50. Available at: https://www.jstage.jst.go.jp/article/ast/39/1/39_E1762/_article (Accessed: 20 February 2020).

Anaconda Software Distribution (2017) Computer software Conda. Available at: https://docs.conda.io (Accessed: 20 March 2020).

Anjos A, El-Shafey L, Wallace R et al. (2012) Bob: a free signal processing and machine learning toolbox for researchers. In: *2012 ACM international conference on Multimedia*. Nara, JP, pp. 1449-1553. Available at: https://publications.idiap.ch/downloads/papers/2012/Anjos_Bob_ACMMM12.pdf (Accessed: 17 February 2020).

Anjos A, Gunther M, de Freitas Pereira T, Korshunov P et al. (2017) Continuously reproducing toolchains in pattern recognition and machine learning experiments. In: *2017 Conference on Neural Information Processing Systems.* Long Beach, USA, pp.1-8. Available at: http://publications.idiap.ch/downloads/papers/2017/Anjos_ICML2017-2_2017.pdf (Accessed: 17 February 2020).

ASVSpoof2019 Consortium (2019) ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan. Available at: http://www.asvspoof.org/asvspoof2019/ asvspoof2019 evaluation plan.pdf (Accessed: 20 February 2020).

Boersma P and Weenink D (2020) Praat: doing phonetics by computer. Available at: http://www.praat.org/ (Accessed: 20 April 2020).

Bryer J and Speerschneider K (2016) *Package 'likert'.* Available at: https://cran.r-project.org/web/packages/likert/likert.pdf (Accessed: 22 February 2020).

Clifford BR (1980) Voice Identification by Human Listeners: On Earwitness Reliability. *Law and Human Behaviour* 4(4): 373-394. Available at: https://www.jstor.org/stable/1393857 (Accessed: 25 February 2020).

Descript (2017) Lyrebird: Ultra-realistic voice cloning and text to speech. Available at: https://www.descript.com/lyrebird-ai?source=lyrebird (Accessed: 12 February 2020).

De Simone F, Goldmann L, Lee J-S et al. (2011) Towards high-efficiency video coding: Subjective evaluation of potential coding technologies. *Journal of Visual Communication and Image Representation* 22(8): 734-748. Available at: https://www.epfl.ch/labs/mmspg/wp-content/uploads/2019/01/spie2010_desimone.pdf (Accessed: 12 February 2020).

Dessimoz D (2004) *Reconnaissance de locuteurs: comparaison de performances entre la reconnaissance auditive par des profanes et de systemes automatiques.* Master Thesis, Université de Lausanne, CH.

Dodson CS and Dobolyi DG (2015) Misinterpreting Eyewitness Expressions of Confidence: The Featural Justification Effect. *Law and Human Behavior* 39(3): 266-280. Available at: https://www.researchgate.net/publication/271334015_Misinterpreting_Eyewitness_Expressions_of_Confidence_The_Featural_Justification_Effect (Accessed: 23 March 2020).

Ergünay SK, Khoury E, Lazaridis A and Marcel S (2015) On the vulnerability of speaker verification to realistic voice spoofing. In: *IEEE 2015 International Conference on Biometrics*: *Theory, Applications and Systems*. Huston, USA, pp. 1-8. Available at: http://publications.idiap.ch/index.php/publications/show/3185 (Accessed: 19 March 2020).

Eriksson A (2010) The Disguised Voice: Imitating Accents or Speech Styles and Impersonating Individuals. In: Llamas C and Watt C (eds) *Languages and Identities*. Edinburgh: Edinburgh University Press, pp. 86-98.

Evans N, Kinnunen T, Yamagishi J et al. (2014) Speaker Recognition Anti-spoofing. In: Marcel S, Nixon MS, and Li SZ (eds) *Handbook of Biometric Anti-Spoofing*. London: Springer, pp. 125–146.

Farrús M (2018) Voice Disguise in Automatic Speaker Recognition. *ACM Computing Surveys* 51(4): 1-22. Available at: https://repositori.upf.edu/bitstream/handle/10230/35866/farrus_ACMComput_voice.pdf?sequence=1&isAllowed=y (Accessed: 14 February 2020).

Hansen JHL and Hasan T (2015) Speaker Recognition by Machines and Humans: A tutorial review. *IEEE Signal Processing Magazine* 32(6): 74-99. Available at:

https://www.researchgate.net/publication/282940395_Speaker_Recognition_by_Machines_and_Humans_A_tutorial_review (Accessed: 04 February 2020).

Harwell D (2019) An artificial-intelligence first: Voice-mimicking software reportedly used in a major theft. *The Washington Post*, 5 September. Available at: https://www.washingtonpost.com/technology/2019/09/04/an-artificial-intelligence-first-voice-mimicking-software-reportedly-used-major-theft/ (Accessed: 12 February 2020).

Hautamäki V, Kinnunen T, Nosratighods M et al. (2010) Approaching Human Listener Accuracy with Modern Speaker Verification. In: *Interspeech 2010 Annual Conference of the International Speech Communication Association*. Makuhari, JP, pp. 1473-1476. Available at: https://www.isca-speech.org/archive/interspeech_2010/i10_1473.html (Accessed: 12 February 2020).

Heinzen E, Lennon R and Hanson A (2020) *Package 'arsenal': comparedf function.* Available at: https://cran.r-project.org/web/packages/arsenal/vignettes/comparedf.html#introduction (Accessed: 22 February 2020).

Hollien H (2002) Forensic Voice Identification. London: Academic Press.

Hossfeld T, Hirth M, Redi J, Mazza F et al. (2014) *Best Practices and Recommendations for Crowdsourced QoE.* Technical report, QUALINET. Available at: https://www.semanticscholar.org/paper/Best-Practices-and-Recommendations-for-Crowdsourced-Hoßfeld-Hirth/b3c7a89362cdc109ff2e7141af67773a4220c55e (Accessed: 12 February 2020).

Huang WC, Wu YC, Kobayashi K, Peng YH et al. (2019) Generalization of Spectrum Differential based Direct Waveform Modification for voice Conversion. In: *2019 ISCA Speech Synthesis Workshop*. Vienna, AU, pp.57-62. Available at: https://arxiv.org/abs/1907.11898 (Accessed: 21 February 2020).

International Organization for Standardization ISO/IEC 30107-1:2016 (2016) Information technology-Biometric presentation attack detection — Part 1: Framework.

International Organization for Standardization ISO/IEC 30107-3:2017 (2017) Information technology-Biometric presentation attack detection — Part 3: testing and reporting.

International Telecommunication Union Radiocommunication Sector (ITU-R) (2019) *Recommendation BS.1284-2 Subjective assessment of sound quality.* Available at: https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1284-2-201901-I!!PDF-E.pdf (Accessed: 17 February 2020).

International Telecommunication Union Radiocommunication Sector (ITU-R) (1990) *Recommendation BS.562-3 Subjective assessment of sound quality.* Available at: https://www.itu.int/rec/R-REC-BS.562-3-199006-W/en (Accessed: 17 February 2020).

Kamble MR, Sailor HB, Patil HA and Li H (2020) Advances in anti-spoofing: from the perspective of ASVspoof challenges. *APSIPA Transactions on Signal and Information Processing* 9(2): pp.1-8. Available at: https://www.researchgate.net/publication/338604760_Advances_in_anti-*spoofing*_from_the_perspective_of_ASVspoof_challenges (Accessed: 14 February 2020).

Kinnunen T, Sahidullah M, Delgado H and Todisco M (2017) The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection. In: *Interspeech 2017 Annual Conference of the International Speech Communication Association*. Stockholm, SE, pp.1-6. Available at: https://www.asvspoof.org/asvspoof2017overview_cameraReady.pdf (Accessed: 02 March 2020).

Korshunov P (2019) Lab-Speaker-Recognition-Jupyter Notebook: *02.features.spectrum.ipynb*. GitLab. Available at: https://gitlab.idiap.ch/biometric-resources/lab-speaker-recognition/blob/master/notebooks/02.features.spectrum.ipynb (Accessed: 14 February 2020).

Korshunov P, Hanhart P, Richter T, Artusi A et al. (2015) Subjective quality assessment database of HDR images compressed with JPEG XT. In: *2015 International Workshop on Multimedia Experience.* Costa Navarino, GR, pp. 1-6. Available at: https://www.researchgate.net/publication/275212302_Subjective_quality_assessment_database_of_HDR_images_compressed_with_JPEG_XT (Accessed: 14 May 2020).

Korshunov P and Marcel S (2016a) Cross-database evaluation of audio-based spoofing detection systems. In: *Interspeech 2016 Annual Conference of the International Speech Communication Association.* San Francisco, USA, pp. 1705-1709. Available at: http://www.isca-speech.org/archive/Interspeech_2016/abstracts/1326.html (Accessed: 16 February 2020).

Korshunov P, Marcel S, Muckenhirn H, Gonçalves AR et al. (2016b) Overview of BTAS 2016 Speaker Anti-*spoofing* Competition. In: *2016 BTAS International Conference on Biometrics Theory, Applications and Systems.* Niagara Falls, USA, pp. 1-6. Available at: https://publications.idiap.ch/downloads/papers/2017/Korshunov_BTAS_2016.pdf (Accessed: 14 May 2020).

Korshunov P, Nemoto H, Skodras A et al. (2014) Crowdsourcing-based evaluation of privacy in HDR images. In: Schelkens P, Ebrahimi T, Cristóbal G et al. (eds) *SPIE 2014 Photonics Europe*. Brussels, BE, pp.1-6. Available at: http://ieeexplore.ieee.org/document/7148119/ (Accessed: 14 February 2020).

Künzel HJ (1994) Current approaches to forensic speaker recognition. In: *ESCA 1994 Workshop on automatic speaker recognition, identification and verification*. Martigny, CH, pp.135-141. Available at: https://www.isca-speech.org/archive_open/archive_papers/asriv94/sr94_135.pdf (Accessed: 12 February 2020).

Jia Y, Zhang Y, Weiss R, Wang Q et al. (2018) Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In: Becker S, Thrun S, Obermayer K (eds) *Advances in Neural Information Processing Systems*. Northampton: Curran Associates, pp. 4480-4490.

Lau YW, Wagner M and Tran D (2004) Vulnerability of speaker verification to voice mimicking. In: *IEEE 2004 international symposium on Intelligent multimedia, video and speech processing*. Paris, FR, pp. 145-148. Available at: https://ieeexplore.ieee.org/abstract/document/1434021 (Accessed: 14 February 2020).

Lindh J and Morrison GS (2011) Humans versus machine: Forensic voice comparison on a small database of Swedish voice recordings. In: *2011 International Congress of Phonetic Sciences.* Hong Kong, CN, pp. 1-4. Available at: https://pdfs.semanticscholar.org/67d2/502ddb50910babc3babbad1517a772a45105.pdf (Accessed: 02 March 2020).

Maganti HK, Motlicek P and Gatica-Perez D (2006) Unsupervised Speech/Non-Speech Detection for Automatic Speech Recognition in Meeting Rooms. In: *IEEE 2007 International Conference on Acoustics, Speech and Signal Processing*. Honolulu, HI, pp. 1037-1040. Available at: https://publications.idiap.ch/downloads/reports/2006/rr06-57.pdf (Accessed: 23 March 2020).

Maher RC (2018) Principles of Forensic Audio Analysis. Modern Acoustics and Signal Processing. Cham: Springer.

Mak MW and Yu HB (2014) A study of voice activity detection techniques for NIST speaker recognition evaluations. *Computer Speech and Language* 28(1): 295-313. Available at: https://www.sciencedirect.com/science/article/pii/S0885230813000533 (Accessed: 23 March 2020).

Mariéthoz J and Bengio S (2006) Can a Professional Imitator Fool a GMM-Based Speaker Verification System? Technical Report, IDIAP.
Available at: http://publications.idiap.ch/index.php/publications/show/356 (Accessed: 14 February 2020).

Masthoff H (2013) A report on a voice disguise experiment. *International Journal of Speech Language and the Law* 3(1): 160-167.
Available at: https://journals.equinoxpub.com/IJSLL/article/view/17245 (Accessed: 14 February 2020).

Mohammadi A (2020) *Trustworthy Face Recognition. Improving Generalization of Deep Face Presentation Attack Detection*. PhD Thesis, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH. Available at: http://publications.idiap.ch/index.php/publications/show/4284 (Accessed: 09 April 2020).

Morrison GS, Sahito FH, Jardine G, Djokic D, et al. (2016) INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International* 263, 92-100. Available at: https://bib-ezproxy.epfl.ch:2061/science/article/pii/S0379073816301311
(Accessed: 09 April 2020).

Muckenhirn H (2019) *Trustworthy speaker recognition with minimal prior knowledge using neural networks*. PhD Thesis, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH.

Nixon KA, Aimale V and Rowe RK (2008) Spoof Detection Schemes. In: Jain A, Flynn P and Ross AA (eds) *Handbook of biometrics*. Boston: Springer, pp. 403-423.

Perrot P and Chollet G (2012) Helping the Forensic Research Institute of the French Gendarmerie to Identify a Suspect in the Presence of Voice Disguise or Voice Forgery. In: Neustein A and Patil HA (eds) *Forensic Speaker Recognition*. New York: Springer, pp.469-503.

Perrot P, Chollet G and Aversano G (2005) Voice Disguise and Automatic Detection: Review and Perspectives. In: *2005 Progress in Nonlinear Speech Processing, Workshop on Nonlinear Speech*. Crete, GR, pp. 101-117. Available at: https://www.researchgate.net/publication/220828901_Voice_Disguise_and_Automatic_Detection_Review_and_Perspectives (Accessed: 12 February 2020).

Pollitt M, Casey E, Jaquet-Chiffelle DO et al. (2018) A Framework for Harmonizing Forensic Science Practices and Digital/Multimedia Evidence: OSAC Task Group on Digital/Multimedia Science. Available at: https://www.nist.gov/news-events/news/2018/01/framework-harmonizing-forensic-science-practices-and-digitalmultimedia (Accessed: 12 February 2020).

R Core Team (2012) R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, AUT. Available at: http://www. R-project.org/ (Accessed: 12 February 2020).

Ramachandra R, Stokkenes M, Mohammadi A, Venkatesh S et al. (2019) Smartphone Multi-modal Biometric Authentication: Database and Evaluation. Available at: https://arxiv.org/pdf/1912.02487.pdf (Accessed: 19 March 2020).

Rodman RD (1998) Speaker recognition of disguised voices: A program for research. In: *1998 Consortium on Speech Technology Conference on Speaker Recognition by Man and Machine: Directions for Forensic Applications*. Ankara, TUR, pp. 9-22. Available at: https://pdfs.semanticscholar.org/4e1e/b942fc9678041d76c4ed0703de9cd1a5189e.pdf?_ga=2.164187832.606201686.1581608674-1374856795.1580733200 (Accessed: 14 February 2020).

Rosenberg A and Ramabhadran B (2017) Bias and statistical significance in evaluating speech synthesis with Mean Opinion Scores. In: *Interspeech 2017 Annual Conference of the International Speech Communication Association.* Stockholm, SE, pp. 3976-3980. Available at: https://www.semanticscholar.org/paper/Bias-and-Statistical-Significance-in-Evaluating-Rosenberg-Ramabhadran/b2b1d01336323f3794f54de26567335aa0bcac46 (Accessed: 26 June 2020).

RStudio Team (2015) *RStudio: Integrated Development for R. RStudio*. Boston, USA. Available at: http://www. R-project.org/ (Accessed: 12 February 2020).

Sahidullah M, Delgado H, Todisco M, Kinnunen T et al. (2019) Introduction to Voice Presentation Attack Detection and Recent Advances. In: Marcel S, Nixon MS, and Li SZ (eds) *Handbook of Biometric Anti-Spoofing*. London: Springer, pp. 321-361.

Sahidullah M, Hanilçi C and Kinnunen T (2015) A Comparison of Features for Synthetic Speech Detection. In: *Interspeech 2015 Annual Conference of the International Speech Communication Association.* Dresden. DE, pp. 2087-2091. Available at: https://erepo.uef.fi/bitstream/handle/123456789/4371/sahidullah_comparison_2015.pdf?sequence=1&isAllowed=y 1580733200 (Accessed: 14 February 2020).

Schmitz C (2012) LimeSurvey: An Open Source survey tool. Available at: http://www.limesurvey.org (Accessed: 23 March 2020).

Schröder M, Charfuelan M, Pammi S and Steiner I (2011) Open source voice creation toolkit for the MARY TTS Platform. In: *Interspeech 2011 Annual Conference of the International Speech Communication Association.* Florence, IT, pp. 3253-3256. Available at: https://www.researchgate.net/publication/221481103_Open_source_voice_creation_toolkit_for_the_MARY_TTS_Platform/citations (Accessed: 04 March 2020).

Schroeter J (2008) Basic Principles of Speech Synthesis. In: Benesty J, Sondhi MM and Huang Y (eds) *Springer Handbook of Speech Processing*. Berlin: Springer, pp. 413-428.

Sithara A, Abraham T and Mathewa D (2018) Study of MFCC and IHC Feature Extraction Methods With Probabilistic Acoustic Models for Speaker Biometric Applications. In: *2018 International Conference on Advances in Computing and Communication*. Kochi, IN, pp. 267-276. Available at: https://www.researchgate.net/publication/329046751_Study_of_MFCC_and_IHC_Feature_Extraction_Methods_With_Probabilistic_Acoustic_Models_for_Speaker_Biometric_Applications (Accessed: 04 April 2020).

Stupp C (2019) Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. *Wall Street Journal*, 30 August. Available at: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 (Accessed: 04 February 2020).

Tekin E, Lin W and Roediger HL (2018) The relationship between confidence and accuracy with verbal and verbal + numeric confidence scales. *Cognitive Research: Principles and Implications* 3(1): 41-49. Available at: https://doi.org/10.1186/s41235-018-0134-3 (Accessed: 23 March 2020).

Todisco M, Delgado H and Evans N (2017) A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients. In: *Odyssey*. Bilbao, SP, pp. 283-290. Available at: https://www.asvspoof.org/papers/CSL_CQCC.pdf (Accessed: 23 March 2020).

Todisco M, Wang X, Sahidullah M, Delgado H et al. (2019) ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection. In: *Interspeech 2019 Annual Conference of the International Speech Communication Association*. Graz, AT, pp. 1008-1012. Available at: http://www.isca-speech.org/archive/Interspeech_2019/abstracts/2249.html (Accessed: 04 February 2020).

Van den Oord A, Dieleman S, Zen H et al. (2016) WaveNet: A Generative Model for Raw Audio. In: *2016 ISCA Speech Synthesis Workshop*. Sunnyvale, California, pp. 1-15.
Available at: http://arxiv.org/abs/1609.03499 (Accessed: 17 February 2020).

Van der Lee C, Gatt A, van Miltenburg E et al. (2019) Best practices for the human evaluation of automatically generated text. In*: 2019 International Conference on Natural Language Generation*. Tokyo, JP, pp. 355-368. Association for Computational Linguistics. Available at: https://www.aclweb.org/anthology/W19-8643.pdf (Accessed: 17 February 2020).

Van Dijk M, Orr R, van der Vloed D and van Leeuwen D (2013) A human benchmark for automatic speaker recognition. In: *1st International Conference Biometric Technologies in Forensic Science*. Nijmegen, NL pp. 39-45. Available at: https://repository.ubn.ru.nl/bitstream/handle/2066/119388/119388.pdf (Accessed: 17 February 2020).

Vestman V, Kinnunen T, Hautamäki RG et al. (2019) Voice Mimicry Attacks Assisted by Automatic Speaker Verification. *Computer Speech and Language* 59(1): 36-54. Available at: https://www.sciencedirect.com/science/article/pii/S0885230818303863 (Accessed: 14 February 2020).

Wang X, Yamaghishi J, Todisco M, Delgado H et al. (2019) The AVS spoof 2019 database. EURECOM. Available at: https://arxiv.org/abs/1911.01601 (Accessed: 14 February 2020).

Watts O, Wu Z and King S (2016) Merlin: An Open Source Neural Network Speech Synthesis System. In: *ISCA 2016 Speech Synthesis Workshop*. Sunnyvale, USA, pp. 202-207.
Available at: https://pdfs.semanticscholar.org/8339/47531a8cd6b79d17003adab58abb00edc0f2.pdf (Accessed: 02 February 2020).

Wenndt SJ and Mitchell RL (2012) Machine recognition vs human recognition of voices. In: *IEEE 2012 International Conference on Acoustics, Speech and Signal Processing*. Kyoto, JP, pp. 4245-4248. Available at: http://ieeexplore.ieee.org/document/6288856/ (Accessed: 12 February 2020).

Wester M, Wu Z and Yamagishi J (2015) Human vs Machine Spoofing Detection on Wideband and Narrowband Data. In: *Interspeech 2015 Annual Conference of the International Speech Communication Association.* Dresden, DE, pp. 2047-2051. Available at: https://www.researchgate.net/publication/279448847_Human_vs_Machine_*Spoofing*_Detection_on_Wideband_and_Narrowband_Data (Accessed: 12 February 2020).

Wu Z, Chng ES and Li H (2012) Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In: *Interspeech 2012 Annual Conference of the International Speech Communication Association.* Portland, USA, pp. 1-4. Available at: https://pdfs.semanticscholar.org/617d/f2f1be497d98c0e255d66eb690 af5a97b259.pdf. (Accessed: 20 February 2020).

Wu Z, De Leon PL, Demiroglu C et al. (2016) Anti-*Spoofing* for Text-Independent Speaker Verification: An Initial Database, Comparison of Countermeasures, and Human Performance. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24(4): 768-783.
Available at: https://ieeexplore.ieee.org/document/7400997 (Accessed: 02 February 2020).

Wu Z, Khodabakhsh A, Demiroglu C, Yamagishi J et al. (2015a) SAS: A speaker verification spoofing database containing diverse attacks. In: *IEEE 2015 International Conference on Acoustics, Speech, and Signal Processing*. South Brisbane, AU, pp. 4440-4444.
Available at: https://ieeexplore.ieee.org/document/7178810 (Accessed: 02 February 2020).

Wu Z, Kinnunen T, Evans N, Yamagishi J et al. (2015b) ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In: *Interspeech 2015 Annual Conference of the*

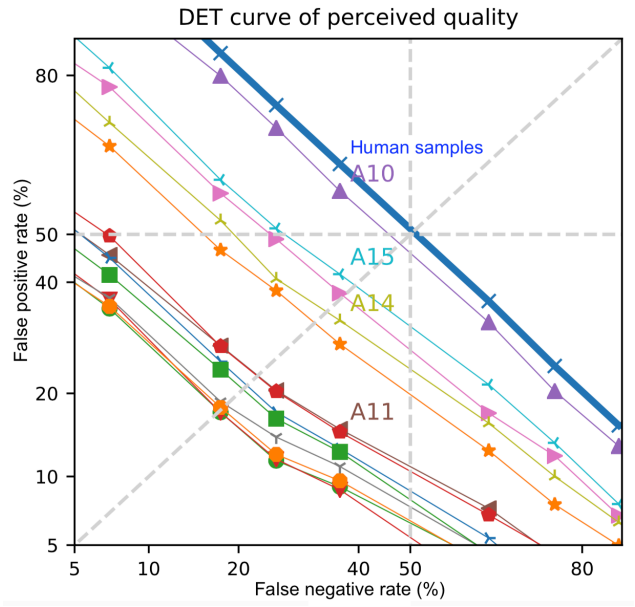
International Speech Communication Association. Dresden, DE, pp. 2037-2041. Available at: https://www.asvspoof.org/is2015_asvspoof.pdf (Accessed: 20 April 2020).

Yamagishi, J, Todisco, M, Sahidullah M et al. (2019) ASVspoof 2019: The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database. University of Edinburgh, The Centre for Speech Technology Research. Available at: https://doi.org/10.7488/ds/2555 (Accessed: 02 February 2020).

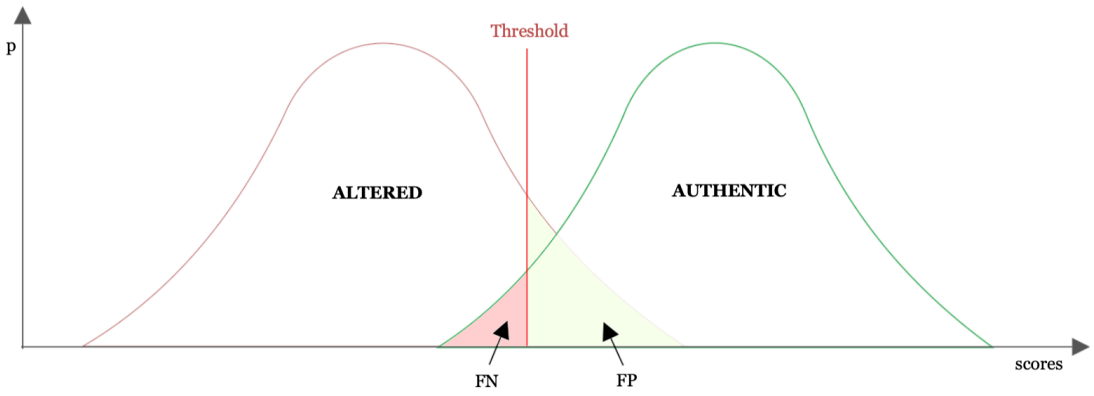
Zetterholm E, Elenius D and Blomberg M (2004) A comparison between human perception and a speaker verification system score of a voice imitation. In: *2004 Australian International Conference on Speech Science and Technology*. Sidney, AU, pp. 393-397. Available at: https://lup.lub.lu.se/search/publication/52907e52-0553-4228-a120-addc5e1f9d24 (Accessed: 14 February 2020).

8 Annex

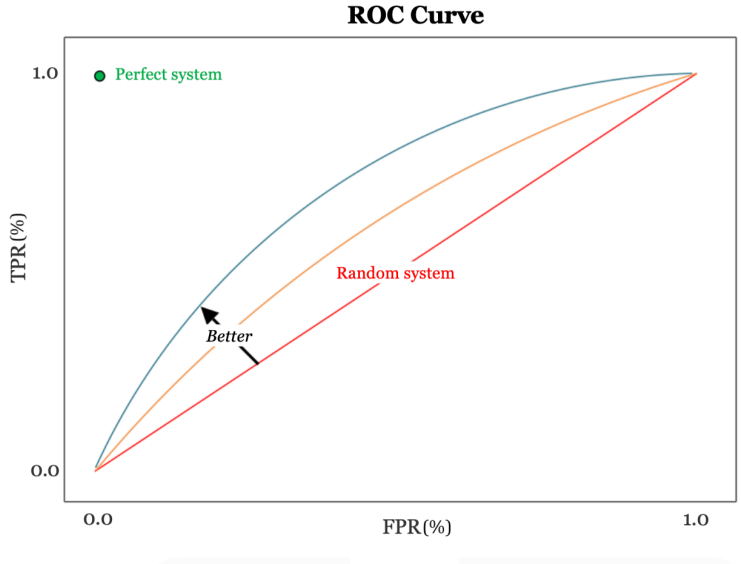
Annex 1: Results as presented in Todisco et al. (2019): the TTS system A10 (purple) is perceived as having a quality level very similar to the one found in human samples (dark-blue) (Figure adapted).



Annex 2: Detection errors and threshold: an illustration.

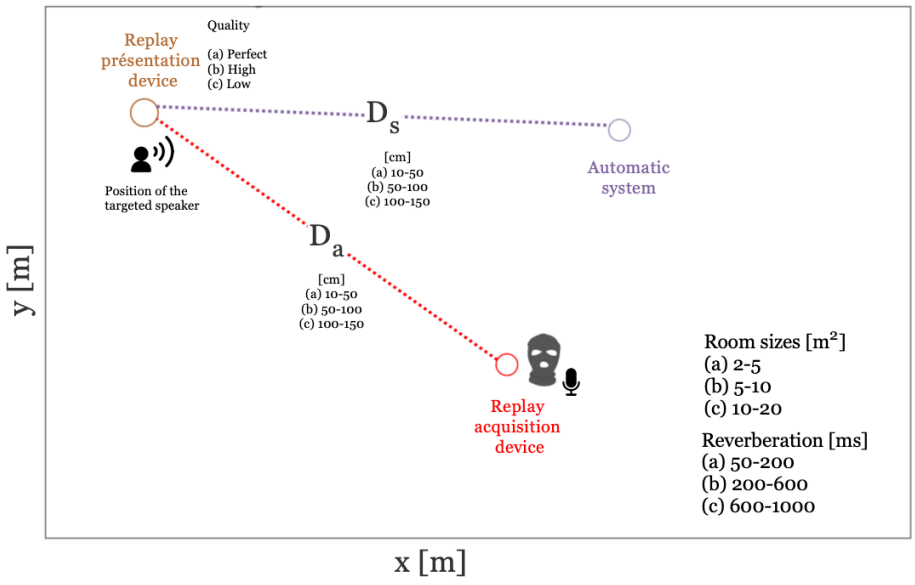


Annex 3: Qualitative representation of a ROC curve. The perfect system has a true positive rate of 1 and a false positive rate of 0 at all thresholds.



Annex 4: Illustration of the PA scenario in ASVspoof2019 (adapted from Wang et al., 2019).

First of all, the simulation of replay disguise considers a distance Da between the acquisition device and the targeted speaker. The recording is then played to the automatic system from a distance Ds using a presentation device producing perfect, high- or low- quality sound. Three room size ranges and three reverberation length ranges are also part of the simulation.



Annex 5: Tables with the information concerning the authentic and altered samples

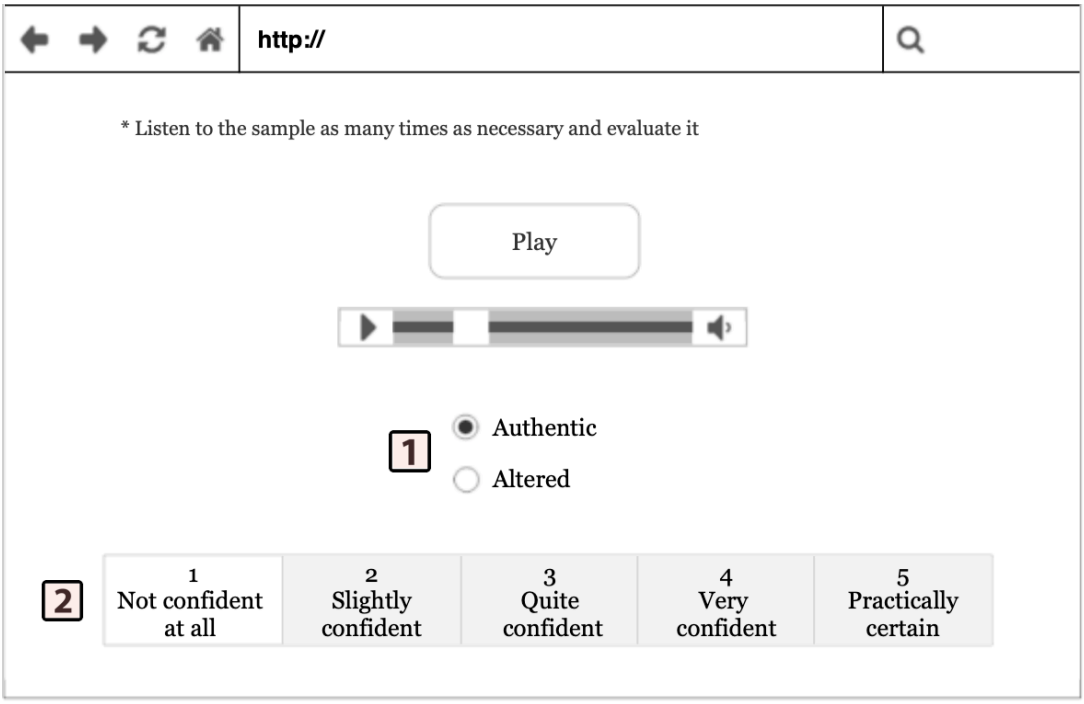
Type	Code	Length (sec)	Transcript
Authentic (R_PA)	PA_E_0011578	1	It was clear
Authentic (R_PA)	PA_E_0077380	3	We are being realistic about the challenges ahead
Authentic (R_PA)	PA_E_0100062	2	First meeting is next week
Authentic (R_PA)	PA_E_0121252	3	The distinction is also supported by doctors
Authentic (R_PA)	PA_E_0102837	3	They will pass it on to Jim Wallis the justice minister
Authentic (R_PA)	PA_E_0123635	2	This could be a recipe for conflict
Authentic (R_PA)	PA_E_0054582	2	It means they actually control the program
Authentic (R_PA)	PA_E_0095729	3	I thought they played very well
Authentic (R_PA)	PA_E_0099310	2	I was inspired by 2 things
Authentic (R_PA)	PA_E_0104476	3	It wasn't an easy decision
Authentic (R_PA)	PA_E_0032774	3	However, we let them back into the game
Authentic (R_PA)	PA_E_0116181	2	He is yet to recieve a reply
Authentic (R_PA)	PA_E_0095558	2	There is a strong envolvement
Authentic (R_PA)	PA_E_0070223	2	So easy does it
Authentic (R_PA)	PA_E_0042468	2	That is the way it is
Authentic (R_PA)	PA_E_0001986	4	The rainbow is a division on white light into many beautiful colors
Authentic (R_PA)	PA_E_0063552	2	It's just not good enough
Authentic (R_PA)	PA_E_0053478	1	Then it will come
Authentic (R_PA)	PA_E_0119879	3	Nobody is pushing out to the right
Authentic (R_PA)	PA_E_0094905	2	I have had a lovely summer
Authentic (R_PA)	PA_E_0114962	2	He will go a long way
Authentic (R_PA)	PA_E_0007451	5	That's just the kind of thing we have
Authentic (R_PA)	PA_E_0065626	4	He has the best performance of the day afterall
Authentic (R_PA)	PA_E_0048678	2	Mr. Smith was dismissive
Authentic (ASVS2019)	LA_E_5849185	4	He has already suffered a good deal of unwanted attention
Authentic (ASVS2019)	LA_E_3757378	3	I need a publishing deal
Authentic (ASVS2019)	LA_E_1027220	2	But he was far from alone
Authentic (ASVS2019)	LA_E_2161075	4	He put some color in scottish history
Authentic (ASVS2019)	LA_E_8739004	2	Your father was a good man
Authentic (ASVS2019)	LA_E_4716734	3	There is a solution, she believes
Authentic (ASVS2019)	LA_E_3593479	2	He was the architect
Authentic (ASVS2019)	LA_E_7769271	2	My whole life has changed
Authentic (ASVS2019)	LA_E_7905661	2	The songs are just so good
Authentic (ASVS2019)	LA_E_2050154	2	He will adress the nation this evening
Authentic (ASVS2019)	LA_E_2291153	3	We just got a phone call on Saturday night
Authentic (ASVS2019)	LA_E_1275973	3	I am a member of the labor party staff
Authentic (ASVS2019)	LA_E_9617894	2	It was still there but right at the end
Authentic (ASVS2019)	LA_E_5313973	1	They can leave at any time
Authentic (ASVS2019)	LA_E_1047198	3	We are really good friends
Authentic (ASVS2019)	LA_E_5432558	1	He has a point
Authentic (ASVS2019)	LA_E_7205247	3	No one has seen this sort of thing before
Authentic (ASVS2019)	LA_E_9578227	2	They thought they couldn't get any
Authentic (ASVS2019)	LA_E_3379472	2	I think she was right
Authentic (ASVS2019)	LA_E_4581379	2	She is a great talent
Authentic (ASVS2019)	LA_E_6314733	2	I said she was very young
Authentic (ASVS2019)	LA_E_3379393	3	It's always nice to play on center court
Authentic (ASVS2019)	LA_E_5323454	4	They final decision was between Scotland and the Republic of Irland
Authentic (ASVS2019)	LA_E_4757272	2	It can be frightening

*RPA = ASVspoof2019_real_PA

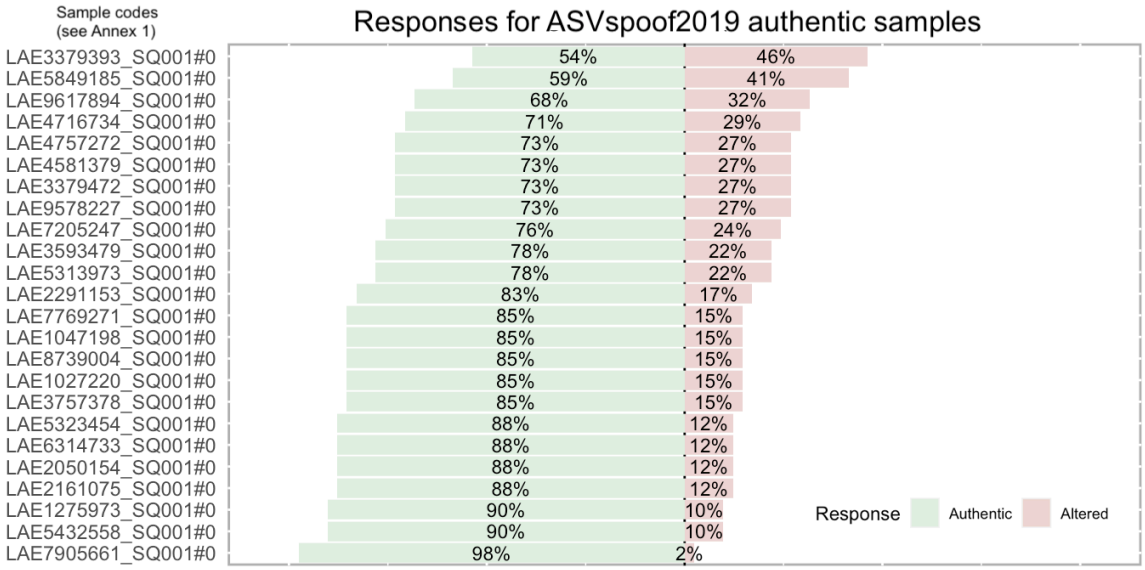
*ASVS2019= ASVspoof2019

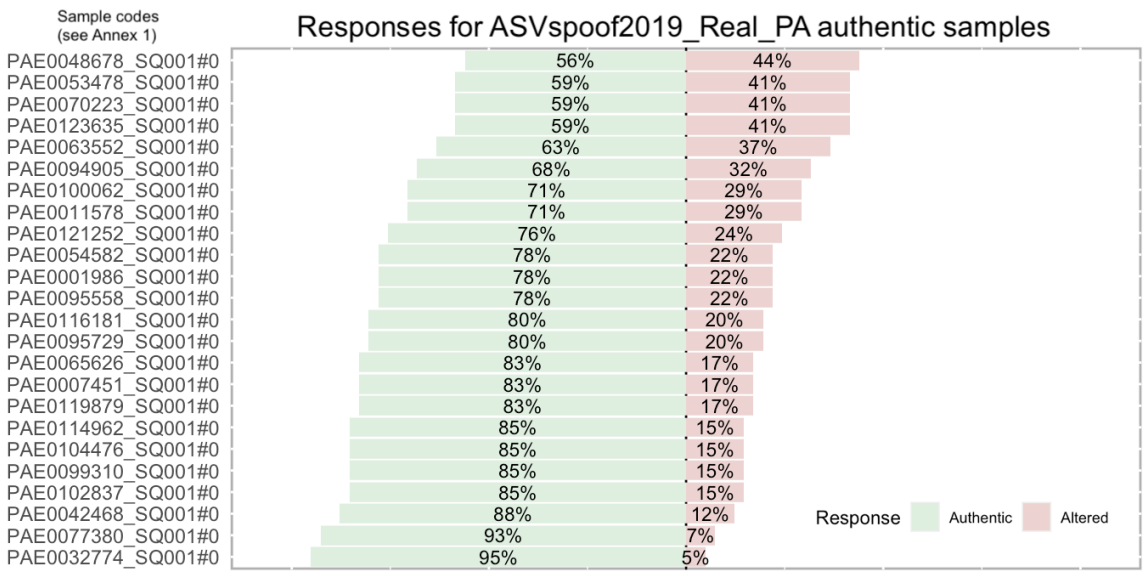
Type	Code	Length (sec)	Transcript
spoof A 10	LA_E_3142969	1	We are not going to forget
spoof A 10	LA_E_6842104	2	The children at the school are all very upset
spoof A 10	LA_E_9977288	2	Our message to the monetary policies commettee is clear
spoof A 10	LA_E_9724819	1	I can lead by example
spoof A 10	LA_E_3170701	2	As usualy there is a variety angle, or is it a trap
spoof A 10	LA_E_4227253	2	This is not the fault of 1 man, of course
spoof A 10	LA_E_4676561	1	I knew staying wasn't an option
spoof A 10	LA_E_3396345	2	But we have built a platform for next season
spoof A12	LA_E_7040813	1	They left me with sadness
spoof A12	LA_E_2729530	2	The jackpot is good for me and for my friends
spoof A12	LA_E_9433024	3	I do not see them being able to do that ever again
spoof A12	LA_E_1210190	2	I don't like the other names they are calling me
spoof A12	LA_E_6374717	1	Nothing has changed
spoof A12	LA_E_8356060	1	We have enough cover
spoof A12	LA_E_3005039	4	The military campaign is a campaign against Osama Binladen
spoof A12	LA_E_1708289	2	The trick is choosing the context
spoof A17	LA_E_2085042	2	It's so awful
spoof A17	LA_E_4361221	4	It looks as thought he will be void
spoof A17	LA_E_9456981	4	Our tasks complete the picture
spoof A17	LA_E_3659898	3	He admitted he was attracted to women
spoof A17	LA_E_5987887	3	This _ is so exciting
spoof A17	LA_E_2475064	2	I will never forget that
spoof A17	LA_E_4860347	3	He was to good for me, to consistent
spoof A17	LA_E_7192618	4	It's fantastic that other women will be able to benefit
spoof PA1	PA_E_0066954	3	He was indeed the grandson of traveling folks
spoof PA1	PA_E_0057813	2	They should not be blamed for it
spoof PA1	PA_E_0052990	2	We have a strong team at the moment
spoof PA1	PA_E_0046753	2	Want to be part of it
spoof PA1	PA_E_0086372	1	It was clear
spoof PA1	PA_E_0079418	3	I did not see anything to begin with
spoof PA1	PA_E_0078036	2	I didn't do it
spoof PA1	PA_E_0086481	2	For me, any manager is good
spoof PA2	PA_E_0039032	2	It should be __
spoof PA2	PA_E_0079613	6	We have____ to go anywhere
spoof PA2	PA_E_0042972	2	I could ease into that
spoof PA2	PA_E_0054663	2	It's an idea
spoof PA2	PA_E_0067222	2	I might come back
spoof PA2	PA_E_0035183	3	He is master deciept and delay
spoof PA2	PA_E_0084133	3	Three other people were threated for minor injuries
spoof PA2	PA_E_0035431	2	It's a do with this place
spoof PA3	PA_E_0059145	3	He's the voice of the survivors
spoof PA3	PA_E_0127340	2	Which is fair enough
spoof PA3	PA_E_0116863	4	Secondly, there are other options for patients
spoof PA3	PA_E_0125677	2	He is not the only one
spoof PA3	PA_E_0074506	2	For once, he was wrong
spoof PA3	PA_E_0095593	2	I sincerely hope not
spoof PA3	PA_E_0040329	2	Is it on the building site?
spoof PA3	PA_E_0065158	3	I have to hand it to the bank

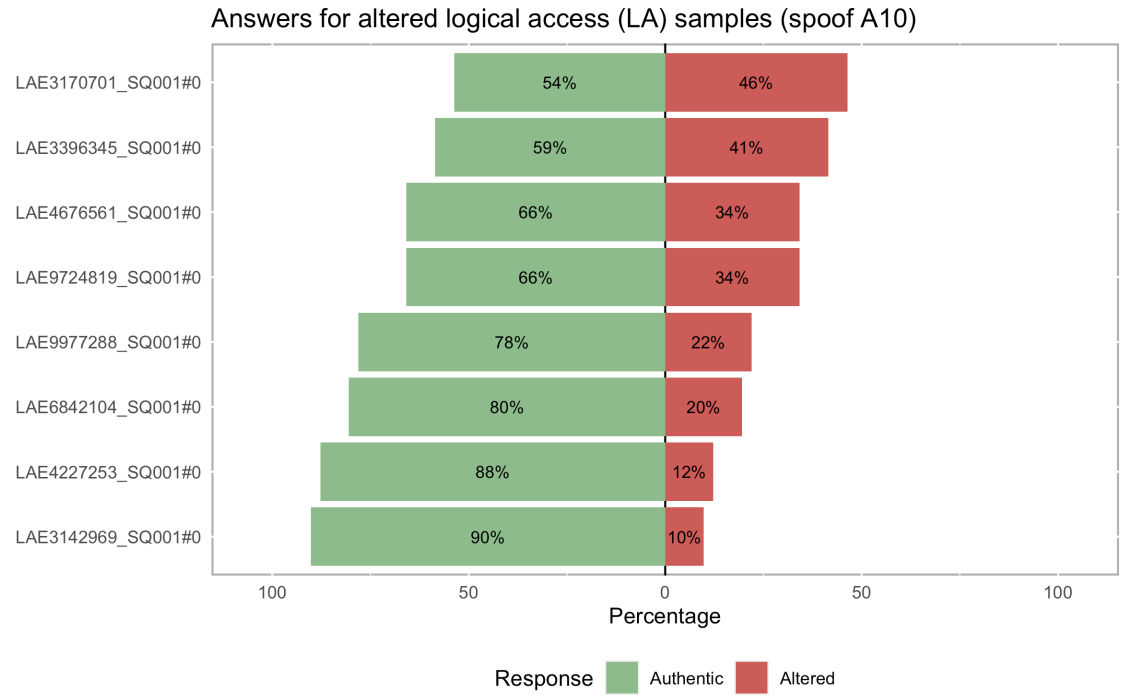
Annex 6: Survey user interface: (1) Forced choice question (2) Confidence level scale

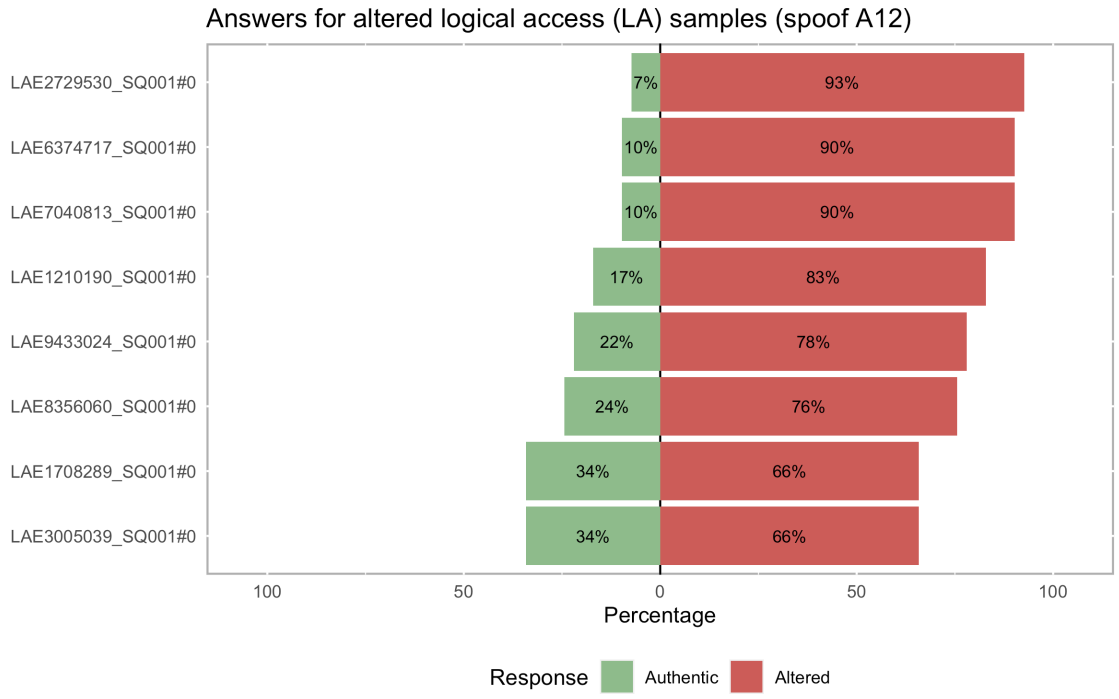


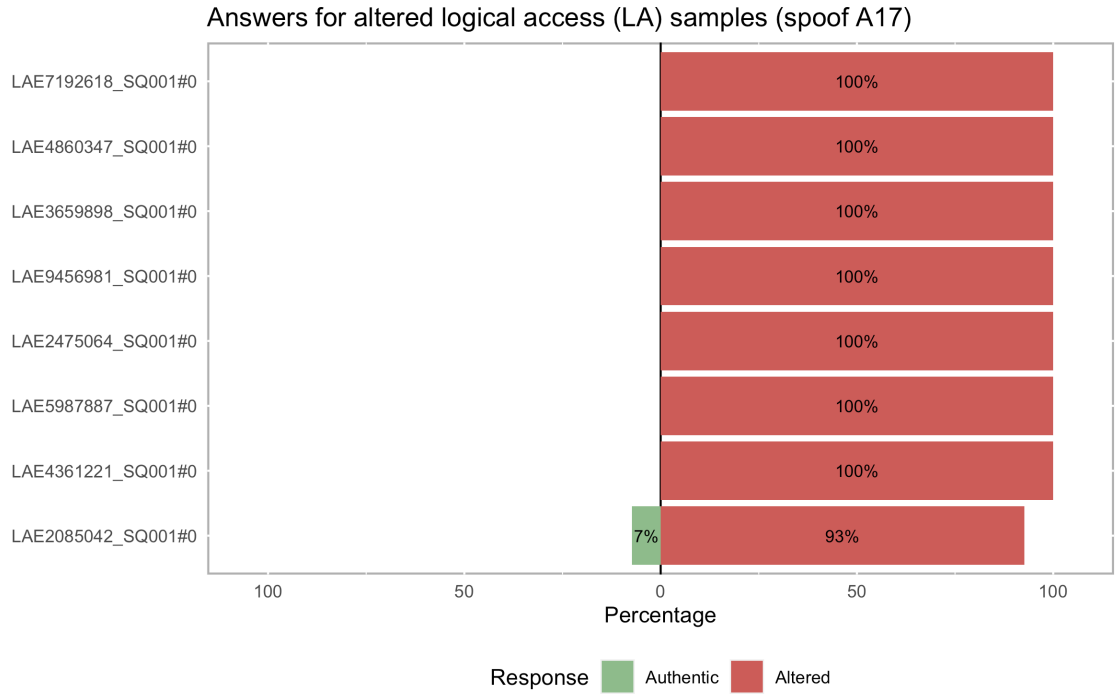
Annex 7: Plots showing the distribution of "authentic" and "altered" answers for each authentic and altered sample. The CRAN *likert* package for Bryer et al. (2016) was used.

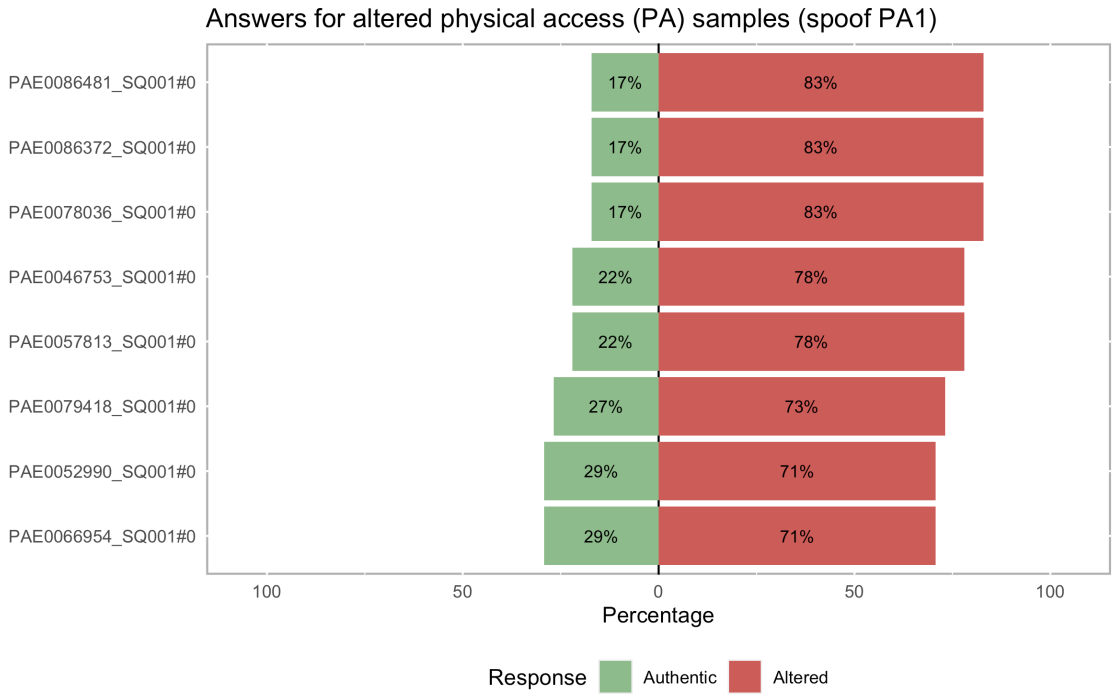


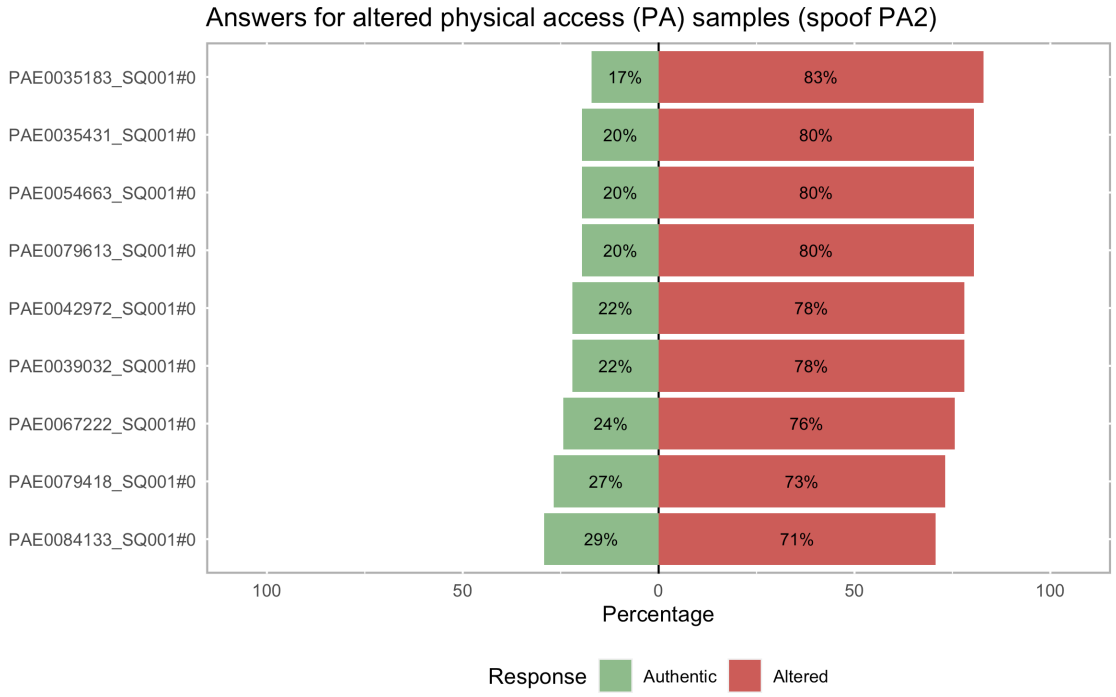

Sample codes
(see Annex 1)
Responses for ASVspoof2019_Real_PA authentic samples
PAE0048678_SQ001#0 56% 44%
PAE0053478_SQ001#0 59% 41%
PAE0070223_SQ001#0 59% 41%
PAE0123635_SQ001#0 59% 41%
PAE0063552_SQ001#0 63% 37%
PAE0094905_SQ001#0 68% 32%
PAE0100062_SQ001#0 71% 29%
PAE0011578_SQ001#0 71% 29%
PAE0121252_SQ001#0 76% 24%
PAE0054582_SQ001#0 78% 22%
PAE0001986_SQ001#0 78% 22%
PAE0095558_SQ001#0 78% 22%
PAE0116181_SQ001#0 80% 20%
PAE0095729_SQ001#0 80% 20%
PAE0065626_SQ001#0 83% 17%
PAE0007451_SQ001#0 83% 17%
PAE0119879_SQ001#0 83% 17%
PAE0114962_SQ001#0 85% 15%
PAE0104476_SQ001#0 85% 15%
PAE0099310_SQ001#0 85% 15%
PAE0102837_SQ001#0 85% 15%
PAE0042468_SQ001#0 88% 12%
PAE0077380_SQ001#0 93% 7%
PAE0032774_SQ001#0 95% 5%
Response Authentic Altered

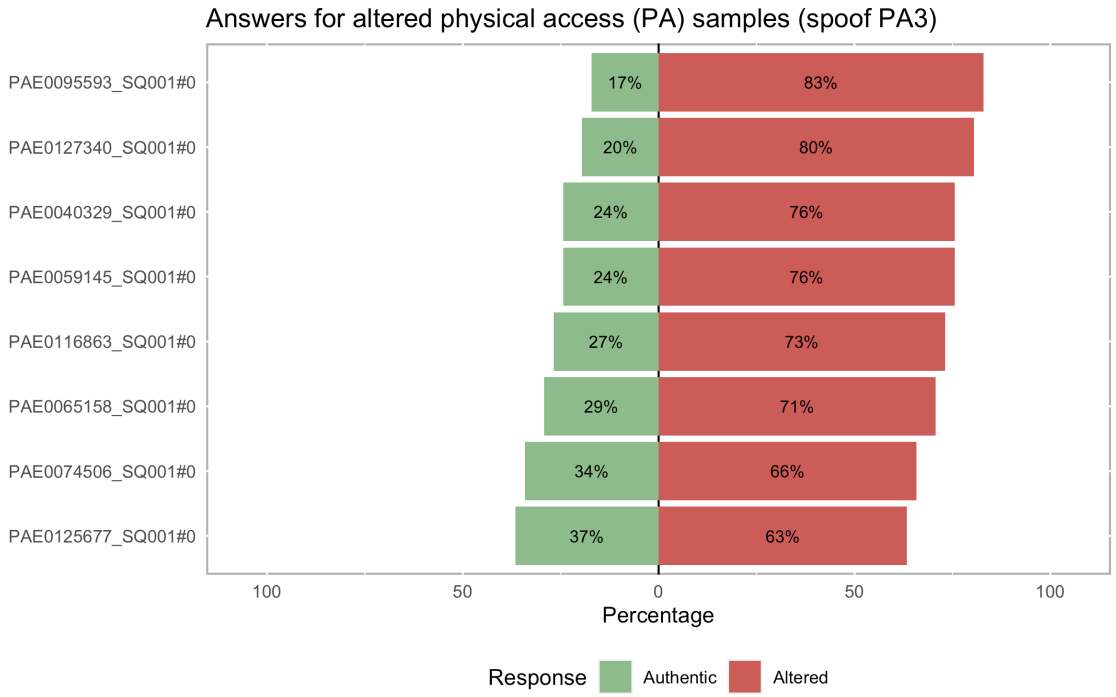

Answers for altered logical access (LA) samples (spoof A10)
LAE3170701_SQ001#0 54% 46%
LAE3396345_SQ001#0 59% 41%
LAE4676561_SQ001#0 66% 34%
LAE9724819_SQ001#0 66% 34%
LAE9977288_SQ001#0 78% 22%
LAE6842104_SQ001#0 80% 20%
LAE4227253_SQ001#0 88% 12%
LAE3142969_SQ001#0 90% 10%
100 50 0 50 100
Percentage
Response Authentic Altered


Answers for altered logical access (LA) samples (spoof A12)
LAE2729530_SQ001#0
LAE6374717_SQ001#0
LAE7040813_SQ001#0
LAE1210190_SQ001#0
LAE9433024_SQ001#0
LAE8356060_SQ001#0
LAE1708289_SQ001#0
LAE3005039_SQ001#0
7%
93%
10%
90%
10%
90%
17%
83%
22%
78%
24%
76%
34%
66%
34%
66%
100
50
0
50
100
Percentage
Response
Authentic
Altered

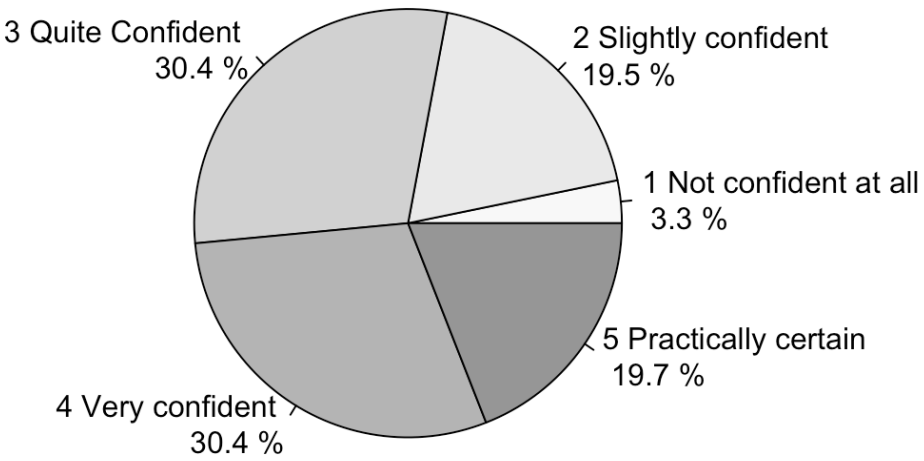

Answers for altered logical access (LA) samples (spoof A17)
LAE7192618_SQ001#0
LAE4860347_SQ001#0
LAE3659898_SQ001#0
LAE9456981_SQ001#0
LAE2475064_SQ001#0
LAE5987887_SQ001#0
LAE4361221_SQ001#0
LAE2085042_SQ001#0
100%
100%
100%
100%
100%
100%
100%
7%
93%
100
50
0
50
100
Percentage
Response
Authentic
Altered


Answers for altered physical access (PA) samples (spoof PA1)
PAE0086481_SQ001#0
PAE0086372_SQ001#0
PAE0078036_SQ001#0
PAE0046753_SQ001#0
PAE0057813_SQ001#0
PAE0079418_SQ001#0
PAE0052990_SQ001#0
PAE0066954_SQ001#0
17%
83%
17%
83%
17%
83%
22%
78%
22%
78%
27%
73%
29%
71%
29%
71%
100
50
0
50
100
Percentage
Response
Authentic
Altered


Answers for altered physical access (PA) samples (spoof PA2)
PAE0035183_SQ001#0
PAE0035431_SQ001#0
PAE0054663_SQ001#0
PAE0079613_SQ001#0
PAE0042972_SQ001#0
PAE0039032_SQ001#0
PAE0067222_SQ001#0
PAE0079418_SQ001#0
PAE0084133_SQ001#0
17%
83%
20%
80%
20%
80%
20%
80%
22%
78%
22%
78%
24%
76%
27%
73%
29%
71%
100
50
0
50
100
Percentage
Response
Authentic
Altered



Annex 8: Pie chart of the reported confidence levels



Annex 9 and 10: Extreme sample analysis for the auditory approach

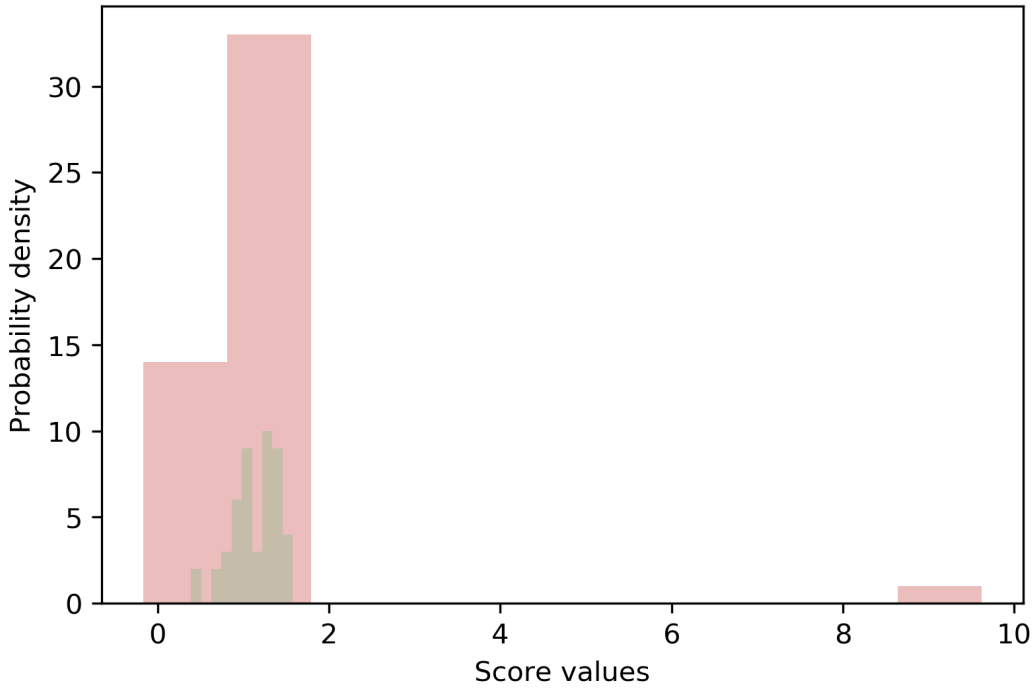
Sample	Mean score	Length(sec)	Speaker gender	Transcript	Analysis
PA_E_0032774	3.28	3	M	However, we let them back into the game	Clear, slow speech.
LA_E_7905661	3.23	2	M	The songs are just so good	Clear, slow speech. Emotional expression trough the intonation.
PA_E_0077380	3.037	3	M	We are being realistic about the challenges ahead	Clear, slow speech. There is no background noise.
LA_E_1275973	2.89	3	M	I am a member of the labor party staff	Clear, slow speech. There is no background noise.
LA_E_2161075	2.70	4	M	He put some color in scottish history	Clear, slow speech. There is no background noise.
...					
LA_E_5849185	0.94	4	M	He has already suffered a good deal of unwanted attention	Very low pitch (F0=77.45 Hz).
PA_E_0048678	0.72	2	M	Mr. Smith was dismissive	"Flat" intonation.
PA_E_0053478	0.70	1	F	Then it will come	Fast speech, slightly raspy voice.
LA_E_3379393	0.48	3	F	It is always nice to play on cinder court	Slight clicking noise at the end of the sentence, absent in most authentic samples.
PA_E_0123635	0.43	2	M	This could be a recipe for conflict	Slightly raspy voice, ever so faint echo.

Highest and lowest scores obtained in authentic samples (highest being the best recognized).

Sample	Mean score	Length(sec)	Speaker gender	Disguise	Transcript	Analysis
LA_E_4227253	2.91	2	F	A10	This is not the fault of one man, of course	Clear, natural speech. There is no background noise.
LA_E_3142969	2.77	1	M	A10	We are not going to forget	Clear, natural speech. There is no background noise.
LA_E_9977288	2.01	2	M	A10	Our message to the monetary policies commettee is clear	Clear speech, a very slight sizzling can be heard.
LA_E_9724819	1.16	1	M	A10	I can lead by example	Clear speech, a very slight sizzling can be heard.
LA_E_4676561	1.01	1	F	A10	I knew staying wasn't an option	Natural speach. There is no background noise but a slight sizzling can be heard.
...						
LA_E_4361221	-4.40	4	M	A17	It looks as thought he will be (?)	Long silence at the beginning. Robotic, unnatural sounding voice. Strong distorsions and bad enunciation.
LA_E_5987887	-4.40	3	F	A17	This (?) is so exciting	Robotic, unnatural sounding voice. Strong distorsions and bad enunciation.
LA_E_2475064	-4.30	2	F	A17	I will never forget that	Robotic, unnatural sounding voice. Abnormally fast speech.
LA_E_9456981	-4.08	4	M	A17	Our tasks complete the picture	Long silence at the beginning. Robotic, unnatural sounding voice. Strong distorsions.
LA_E_3659898	-4.03	3	M	A17	He admitted he was attracted to women	Robotic, unnatural sounding voice. Strong distorsions and bad enunciation.

Highest and lowest scores obtained in altered samples (lowest being the best recognized).

Annex 11: Scores histogram for the LFCC-GMM- based system (AVspoof+Swan) before the trim.silence VAD filter. At a score of 9.82 the outlier LA_E_3005039.



Annex 12 and 13: Extreme sample analysis for the automatic approach (AVspoof+Swan)

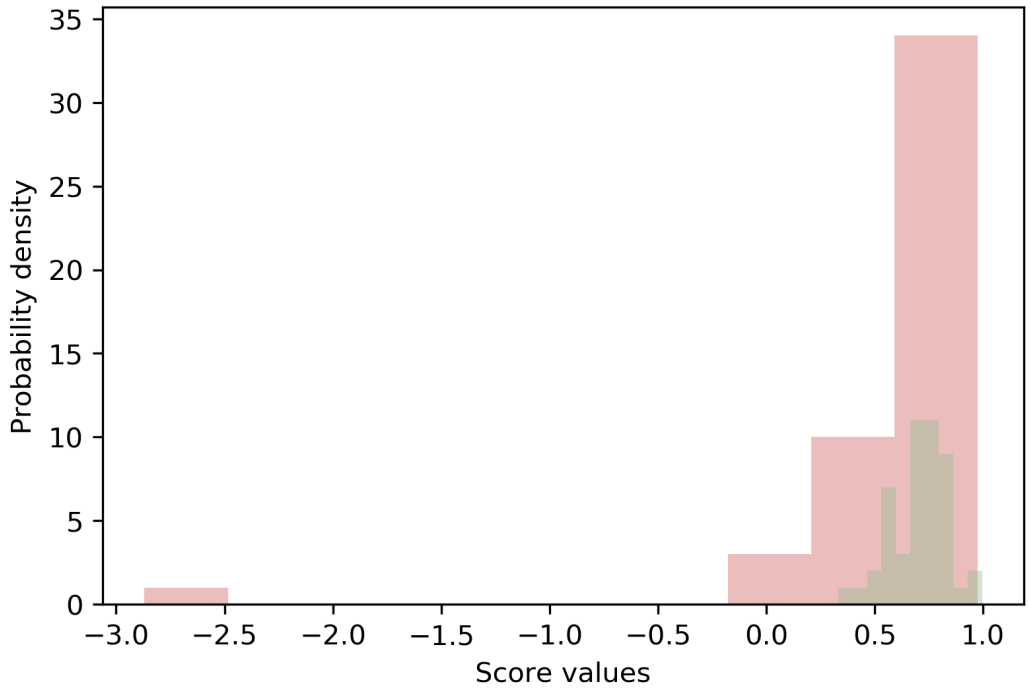
Sample	Score	Length(sec)	Speaker gender	Transcript	Analysis
LA_E_2050154	2.14	2	M	He will adress the nation this evening	Clear, slow speech. There is no background noise.
LA_E_5313973	2.10	1	F	They can leave at any time	Clear speech. There is no background noise.
PA_E_0042468	2.01	2	F	That is the way it is	Clear speech, slightly raspy voice.
LA_E_3379472	1.95	2	F	I think she was right	Clear speech. There is no background noise.
LA_E_3379393	1.94	3	F	It's always nice to play on cinder court	Slight clicking noise at the end of the sentence, absent in most authentic samples.
...					
LA_E_7769271	0.93	2	M	My whole life has changed	Clear, slow speech. Slight background noise.
LA_E_2161075	0.79	4	M	He put some color in scottish history	Clear, slow speech. There is no background noise.
PA_E_0007451	0.77	5	F	That's just the kind of thing we have	Clear, fast speech. Long silence at the end.
LA_E_8739004	0.33	2	M	Your father was a good man	Clear, slow speech. Slight background noise.
PA_E_0114962	-0.10	2	F	He will go a long way	Clear, slow speech. There is no background noise.

Highest and lowest scores obtained in authentic samples (highest being the best recognized.

Sample	Score	Length(sec)	Speaker gender	Disguise	Transcript	Analysis
LA_E_2085042	2.16	2	M	A17	It's so awful	Long silence at the beginning.Robotic, unnatural sounding voice. Strong distorsions.
LA_E_1708289	2.01	2	F	A12	The trick is choosing the context	Clear speech, but a sizzling noise can be heard.
LA_E_3142969	1.92	1	M	A10	We are not going to forget	Clear, natural speech. There is no background noise.
LA_E_2729530	1.81	2	M	A12	The jackpot is good for me and for my friends	Robotic, unnatural sounding voice with distorsions.
LA_E_1210190	1.79	2	M	A12	I don't like the other names they are calling me	Robotic, unnatural sounding voice with distorsions.
...						
LA_E_8356060	0.73	1	F	A12	We have enough cover	Robotic, unnatural sounding voice with distorsions.
LA_E_7192618	0.70	4	F	A17	It's fantastic that other women will be able to benefit	Robotic, unnatural sounding voice. Strong distorsions.
LA_E_5987887	0.58	3	F	A17	This (?) is so exciting	Robotic, unnatural sounding voice. Strong distorsions and bad enunciation.
LA_E_3659898	-0.18	3	M	A17	He admitted he was attracted to women	Robotic, unnatural sounding voice. Strong distorsions and bad enunciation.
LA_E_4361221	-1.38	4	M	A17	It looks as thought he will be void	Long silence at the beginning. Robotic, unnatural sounding voice. Strong distorsions.

Highest and lowest scores obtained in altered samples (lowest being the best recognized).

Annex 14: Scores histogram for the LFCC-GMM- based system (ASVspoof 2015) before the trim.silence VAD filter. At a score of -2.12 the outlier LA_E_3005039.



Annex 15 and 16: Extreme samples analysis for the automatic approach (ASVspoof 2015)

Sample	Score	Length(sec)	Speaker gender	Transcript	Analysis
PA_E_0042468	1.37	2	F	That is the way it is	Clear speech, slightly raspy voice.
LA_E_5432558	1.28	1	F	He has a point	Clear, fast speech.
PA_E_0095558	1.27	2	F	There is a strong involvement	Clear, slow speech.
LA_E_3757378	1.23	3	M	I need a publishing deal	Clear, slow speech.
PA_E_0070223	1.19	2	F	So easy does it	Clear, slow speech.
...					
PA_E_0102837	0.55	3	M	They will pass it on to Jim Wallis the justice minister	Clear, deep voice.
PA_E_0063552	0.38	2	F	It's just not good enough	Clear speech, rustling sound at the end.
PA_E_0065626	0.38	4	F	He has the best performance of the day afterall	Clear, fast speech.
LA_E_7205247	0.28	3	F	No one has seen this sort of thing before	Clear but very monotone speech.
PA_E_0053478	0.23	1	F	Then it will come	Clear, fast speech. Slightly raspy voice.

Highest and lowest scores obtained in authentic samples (highest being the best recognized)

Sample	Score	Length(sec)	Speaker gender	Disguise	Transcript	Analysis
LA_E_6374717	1.08	1	F	A12	Nothing has changed	Quite clear voice, robotic undertones.
LA_E_6842104	1.03	2	M	A10	The children at the school are all very upset	Natural speech. There is no background noise but a slight sizzling can be heard.
PA_E_0067222	1.03	2	F	PA2	I might come back	Natural speech but an echo and background noise can be heard.
LA_E_4676561	0.97	1	F	A10	I knew staying wasn't an option	Natural speech. There is no background noise but a slight sizzling can be heard.
PA_E_0065158	0.97	3	F	PA3	I have to head into the bank	Natural speech but an echo and background noise can be heard.
...						
LA_E_2085042	0.01	2	M	A17	It's so awful	Long silence at the beginning. Robotic, unnatural sounding voice. Strong distorsions.
LA_E_5987887	-0.08	3	F	A17	This (?) is so exciting	Robotic, unnatural sounding voice. Strong distorsions and bad enunciation.
LA_E_4860347.	-0.19	3	F	A17	He was to good for me, to consistent	Long silence at the beginning. Robotic, unnatural sounding voice. Strong distorsions.
LA_E_2475064.	-0.20	2	F	A17	I will never forget that	Long silence at the beginning. Robotic, unnatural sounding voice. Strong distorsions.
LA_E_4361221	-0.97	4	M	A17	It looks as thought he will be void	Long silence at the beginning. Robotic, unnatural sounding voice. Strong distorsions.

Highest and lowest scores obtained in altered samples (lowest being the best recognized)

Annex 17: Digital files

- Rmarkdown code
- Evaluation databases
- Survey files and results
- Automatic system configuration and execution files