Short-Range Audio Channels Security: Survey of Mechanisms, Applications, and Research Challenges

Maurantonio Caprolu, Savio Sciancalepore, Roberto Di Pietro
Division of Information and Computing Technology (ICT)
College of Science and Engineering (CSE), Hamad Bin Khalifa University (HBKU)
Doha, Qatar
mcaprolu@mail.hbku.edu.qa, ssciancalepore@hbku.edu.qa, rdipietro@hbku.edu.qa

***Abstract*— Short-range audio channels have appealing distinguishing characteristics: ease of use, low deployment costs, and easy to tune frequencies, to cite a few. Moreover, thanks to their seamless adaptability to the security context, many techniques and tools based on audio signals have been recently proposed. However, while the most promising solutions are turning into valuable commercial products, acoustic channels are also increasingly used to launch attacks against systems and devices, leading to security concerns that could thwart their adoption. To provide a rigorous, scientific, security-oriented review of the field, in this paper we survey and classify methods, applications, and use-cases rooted on short-range audio channels for the provisioning of security services—including Two-Factor Authentication techniques, pairing solutions, device authorization strategies, defense methodologies, and attack schemes. Moreover, we also point out the strengths and weaknesses deriving from the use of short-range audio channels. Finally, we provide open research issues in the context of short-range audio channels security, calling for contributions from both academia and industry.**

***Index Terms*—Audio Channel Security, Audio-based Authentication, Pairing via Audio, Audio Attacks, Audio for Cyber-Physical Systems Security.**

I. Introduction

Sound, including human speech, is commonly considered as a natural and intuitive means to quickly interact with automatic devices [1]. Indeed, ambient sounds, as well as voice commands issued towards audio-enabled devices, are often conceived as a natural, intuitive, and minimal effort approach for humans to communicate with machines, especially if compared to human-imperceptible Radio Frequency (RF) transmissions and distracting visual-tactile interfaces [2].

Recently, communications using short-range audio channels have attracted increasing interest, in academia as well as industry, as demonstrated by the expected 31.80 USD billions for this specific market expected by the year 2023 [3]. In this context, security-oriented applications of the short-range audio stem in a prominent position. Besides aspects related to the enhanced usability, there are mainly two dominating research directions. For the defense of systems and devices, the general opinion in the research area is that acoustic channels are a more secure communication channel when compared to legacy RF communications. For instance, given that audio channels can be perceived by participating entities, they are usually assumed to be robust against active attacks, since any malicious signal could be heard by the participants as well—and hence the attack could be detected [4]. At the same time, the increasing diffusion and affordability of Machine Learning (ML) techniques have boosted the efficiency of audio signals classification and identification, leading to low-cost and disrupting attacks. Indeed, recent efforts demonstrated that the sound emitted from specific devices, such as keyboards and 3D printers, leaks unique information about the specific performed task, e.g., the pressed button and the printed object features, respectively [5], [6]. Considering that these devices do not provide any inherent defense strategy to protect information leaked on the acoustic channel, such attacks are stealthy and potentially disrupting [7].

Inspired by these research challenges and by the relevant contributions in the area over the last years, in this paper we survey the mechanisms, applications, use-cases and research challenges involving the use of short-range audio channels for the provision of security services. We show how the peculiar features of the short-range acoustic signals, including physical proximity between communicating entities and audibility of the communications, are used to provide additional authentication schemes, pairing, context-based authorization, and defense tools, as well as to launch attacks against systems and devices. As a novel contribution, we divide and classify the scientific literature according to the provided security feature. In addition, for each of the categories identified in our survey, we explain and compare the most important contributions in the actual literature. Furthermore, we also identify some crucial research challenges, whose further investigation could unleash the full potential of audio-based security solutions, and pave the way to their large-scale adoption.

Overall, the following key contributions are provided in this manuscript (see Fig. 1 for a graphical overview).

- We survey the mechanisms, applications, use-cases and research challenges involving the use of short-range audio channels for the provision of security services to systems and users.
- We classify the security services provided via short-range audio channels into five (5) different categories, including authentication schemes (Section III), pairing (Section IV),

This is a personal copy of the authors. Not for redistribution. The final version of the paper is available through the IEEExplore Digital Library, at the link: https://ieeexplore.ieee.org/document/8967104, with the DOI: 10.1109/COMST.2020.2969030.

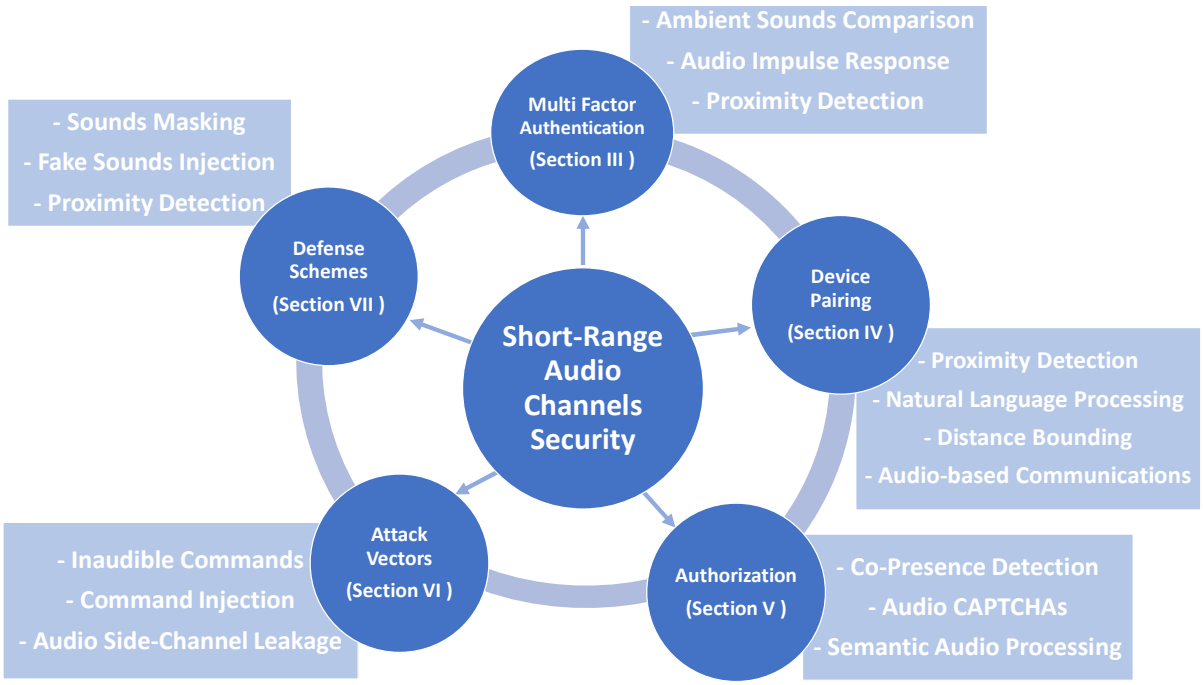

Fig. 1. Overview of the taxonomy and the organization of our paper. We classify security schemes leveraging the audio channel based on the specific security services they intend to provide, identifying two-factor authentication approaches (Section III), secure device pairing strategies (Section IV), authorization techniques (Section IV-B), attack vectors (Section VI), and systems defense schemes (Section VII).

context-based authorization (Section IV-B), attacks (Section VI), and defense schemes (Section VII), detailing and cross-comparing them along the most important features.
- We identify several important research challenges in each of the aforementioned domains, where further scientific contributions are still required to fill the actual gaps (Section VIII).
- We identify a few promising future research directions involving the usage of short-range audio channels, whose thorough investigation could further unleash the potential of these communication technologies (Section VIII).

Paper	Main Topic Addressed
[8]	Survey of the approaches supporting the pairing between devices leveraging the physical context shared by participating entities.
[9]	Survey of strategies used to establish the co-presence between communicating devices.
[10]	Survey of multi-factor authentication techniques.
[11]	Survey of attacks exploiting side-channels and unintentional information leakage.
[12]	Survey of audio and visual CAPTCHAs used to distinguish humans from bots.
[13]	Survey of steganography and other information hiding techniques.

TABLE I
OVERVIEW OF RELATED SURVEYS AND THEIR RESPECTIVE TOPICS

As highlighted in Tab. I, some security solutions based on audio channels have been quickly touched in the recent literature by few surveys, focusing either on context-based security [8], [9], authentication solutions [10], side-channel attacks [11], CAPTCHAs [12] or forensic applications [13]. However, to the best of our knowledge, our contribution is the first to specifically survey, classify, and systematize security methods, applications, and use-cases using only short-range audio channels, as well as to evaluate such strategies across different use-cases and objectives, and to propose further research challenges—specific to this domain.

Being focused on short-range audio channels, our survey does not include underwater communication schemes, audio watermarking, and audio forensic techniques, as they involve either long distances or considerations related to the audio signal processing, not specifically meant for secure communication and applications.

We believe that this work could attract researchers coming from heterogeneous backgrounds, either interested in finding and applying unconventional security solutions to classical security issues, or researchers particularly skilled in the specific audio domain, finding in security applications the perfect area where they could provide different perspectives and innovative solutions. Indeed, through the lenses of scientifically rigorous work, the interested readers will find meaningful mechanisms and solutions related to the physical properties of audio channels used for security applications, and how they are integrated into modern computer systems and devices to enhance their security level. Further, open research challenges are also highlighted, to inspire further research in this exciting domain.

The rest of this paper is organized as follows: Section II provides the necessary preliminary details about the audio processing, as well as the most important techniques used in the literature for security services via audio. Section III focuses on applications of audio to Two-Factor Authentication protocols, Section IV provides a comprehensive overview

of the techniques based on audio used for the pairing of previously unknown devices, while Section IV-B details how audio can be used to authorize the usage of systems and devices. The most important attacks exploiting audio channels and their components are discussed in Section VI, while the use of audio as a defense tool is discussed in Section VII. The most important research challenges arising from the above discussions are summarized and explained in Section VIII, while Section IX closes the paper.

II. Background

In this section, we briefly introduce some technical concepts related to the analysis and usage of short-range audio signals. While a thorough analysis of these concepts is out of the scope of this contribution, their high-level description will be useful to fully catch differences and similarities between scientific contributions applying audio signals for different security-oriented tasks. Section II-A illustrates the basic features of audio signals and their processing on modern computers, while the fundamental phenomena behind human speech generation and electronic speech reproduction are presented in Section II-B and Section II-C, respectively. Section II-D provides an overview of the most used techniques to compare audio signals is provided, while the logic of the masking sound technique and the *cocktail party problem* are provided in Section II-E and Section II-F. Finally, Section II-G highlights how short-range audio links can be used to estimate the distance between communicating entities.

A. Analyzing Audio Signals

The audible frequencies range, i.e., the range of audio signals that could be heard by human ears, spans the bandwidth [20 - 20,000] Hz. In practice, these limits fluctuate a few Hz above or below according to the particular individual, as specific features of the human ear can slightly increase or decrease sounds perception. The sounds below the minimum audible frequency are defined infra-sounds, while sounds above the 20,000 Hz threshold are ultrasounds.

A common representation for the spectral power density in such a wide bandwidth is based on the use of octave bands, splitting the above range into 11 non-overlapping sub-bands. Analyzing audio signals in octave bands (or in their sub-components) allows for a more precise investigation of the features of the complex signal, as well as an in-depth analysis of the Signal to Noise Ratio (SNR) across the audible range [14].

Specifically, octave bands are selected such that the center frequency of a given band is twice the center frequency of the immediately lower band. Fig. 2 shows the details of the octave bands, reported according to their calculated center frequencies.

Note that the lower and upper limits of the audible range are essentially defined by the octave bands containing the lower and the higher frequencies that the human ear can receive.

Octave bands are mainly used when first analyzing a sound, to identify the portion of the audio spectrum where the audio signal is mainly concentrated.

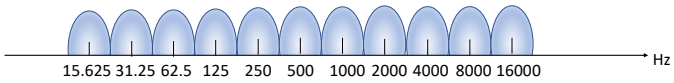

Fig. 2. Octave bands used for audio signal representation and analysis in digital systems. There are 11 octave bands, and the center of any octave band is chosen such that it has a frequency that is exactly the double of the center frequency of the previous octave band.

In addition, octave bands are usually further divided into three ranges, namely *one-third-octaves* bands. The use of one-third-octave bands has been standardized as a reference baseline for use in commodity scientific instruments and measurements. They are mainly used in environmental and noise control applications, to provide a further in-depth outlook on noise levels across the frequencies.

B. Human Speech sounds

The human speech is the result of complex interactions between different elements of the human phonetic apparatus, including the vocal cords, teeth, lips, and mouth structure, to name a few, as shown in Fig. 3.

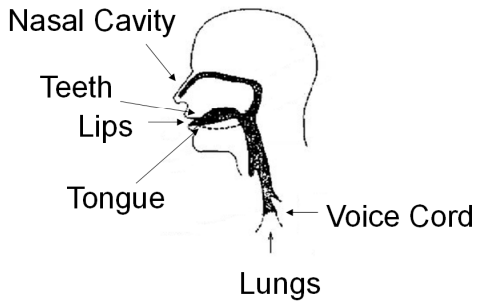

Fig. 3. Physical elements of the human phonetic apparatus involved in speech generation. The vocal cords generate the sound, characterized by a minimum frequency, i.e., the fundamental frequency, and multiple harmonics at higher frequencies. The sound is shaped by the tongue, the lips, the teeth, and the nasal cavity, that filter out some harmonics and produce the final sound emitted by the lips.

Without loss of generality, a simplistic model is frequently used, where the speech sound is modeled as the combination of two different effects. A voiced source, physically coincident with the vibration of the vocal cords, generates the signal, characterized by a minor frequency, namely the *Fundamental Frequency*, and harmonic frequencies at multiples of the fundamental one. This signal is further shaped by a filter, modeling the combined effect of the tongue, lips, teeth and nasal cavity. Overall, these components remove higher-frequencies harmonics and produce the final signal coming out from the lips. Typically, women and men have fundamental frequencies in the range [165-255] Hz and [85-180] Hz, respectively [15]. This is caused by the structure of the vocal cords, that men have typically larger than women. However, given that no new frequencies are introduced by the elements in the mouth, the voice of an individual is typically uniquely associated with its fundamental frequency, i.e., the lowest frequency in the voice signal, also known as the *pitch* of the speaker's voice. It is worth noting that the final sound produced by a human

is given by the interaction between the action of the lips and the nose, further shaping the tone of the voice.

A technique that is frequently used for speech recognition is the Mel-Frequency Cepstral Coefficient (MFCC), described in [16]. This technique best describes how the human phonetic system perceives sounds. The MFCC technique divides the whole sampled window in very short windows, where the audio signal can be safely assumed as stationary. Then, the Fast Fourier Transform (FFT) of each window is computed and it is shifted to an alternative frequency axis, namely the non-linear Mel scale, further divided into a set of bands, known as the Mel bands. This step is useful given that the speech spectrum is very wide and it does not follow a linear scale. Thus, to represent the speech signal, an alternative logarithmic scale is used, summarizing the spectrum in an increasingly wide bandwidth.

For each Mel band, the Discrete Cosine Transform (DCT) is applied to the logarithm of the power spectral density of the input signal, to provide the final MFCC representation. There are studies, such as [17], that demonstrated the suitability of the MFCC representation for speech recognition, as it models particularly well the variability of the human speech across different users. In addition, the MFCC technique enables further applications, such as the speaker emotion recognition [18].

C. Electronic Speakers Sounds

The simplified internal structure of a modern sound-generation speaker is depicted in Fig. 4.

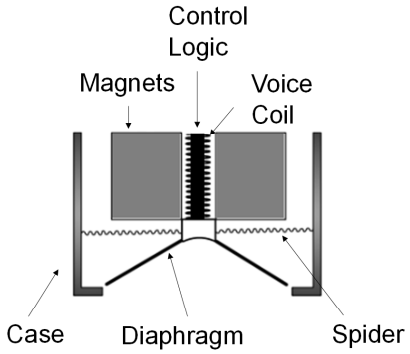

Fig. 4. Internal components of a modern speaker. The control logic provides current to the voice coil, generating a magnetic field. The magnetic field induces the movements of the diaphragm, which modulates air pressure and produces the sound waves.

Modern speakers reproduce sounds through the controlled oscillation of the diaphragm, steered by a control circuit and a voice coil. The control circuit receives an electronic signal from the device where the speaker is connected. The control circuit provides current to the voice coil, which further generates a magnetic field. The voice coil is located in-between two magnets, and the interaction among them causes a force that induces the diaphragm to move. The speaker also contains a spider, physically connecting the diaphragm to the case. Its role is mainly structural, as it should ensure the persistent connection between the diaphragm and the case while also allowing a free oscillation of the diaphragm and the voice coil to return in its original position. Finally, it is worth noting that the material the case consists of is crucial to not deteriorate the produced sound too much. Indeed, any material has its resonance frequency, causing resonating effects in the sub-bass region, in the frequency band [20-80] Hz.

The physical features of the speakers directly impact the performance of the device. Typically, this is expressed via the Frequency Response Curve, indicating the quality of the sounds reproduced by the speaker at a given frequency. Usually, to guarantee optimal performances, speakers are designed to work well with either low frequencies or higher frequencies, due to physical trade-offs (size of the materials). Indeed, to induce higher frequency sounds, the diaphragm should move more air, and hence, the voice coil would need to induce a stronger magnetic field by maintaining the same form factor of the speaker. Thus, it could be possible only if the voice coil is larger, i.e., by changing the physical components in the speaker. This is the reason at the basis of the choice operated by modern sound systems, that are equipped with distinct speakers, each optimized for the bandwidth of the sound to be reproduced [19].

D. Comparing Audio Recordings

Comparing audio recordings to establish their similarity is a classical problem in audio processing, very frequently faced by the approaches available in the literature. Hence, various techniques have been proposed.

A common metric to assess the similarity between the two sound registrations is established via a *similarity score*. It is computed as the average of the maximum normalized cross-correlation over the pair of signal components acquired by the two devices. This metric is particularly useful when the two audio recordings are quasi-synchronized, and an indication of their similarity can be extracted from their time-domain representation. Defining the two signals x and y as n-points discrete time series, namely $x(n)$ and $y(n)$, the cross-correlation $c_{x,y}(l)$ is obtained with reference to a time lag $l \in \{0, 1\}$, as depicted in the following Eq. 1.

$$c_{x,y}(l) = \sum_{i=0}^{n-1} x(i) \cdot y(i-l), \tag{1}$$

with $y(i) = 0$ when $i < 0$ or $i > n-1$. Usually, two audio recordings are characterized by different amplitudes, as the receiving microphones are located at different distances from the emitting sources. To eliminate dependency from the amplitudes, a normalized measure is taken as in the following Eq. 2, to map the correlation values in the interval [-1,1].

$$c'_{x,y} = \frac{c_{x,y}(l)}{\sqrt{c_{x,x}(0) \cdot c_{y,y}(0)}}. \tag{2}$$

Indeed, a value $c'_{x,y} = 1$ indicates that the two recordings have equal evolution in the time domain, independently from their amplitudes. A value $c'_{x,y} = -1$ indicates that they have equal evolution in the time domain but opposite signs, while the value $c'_{x,y} = 0$ indicates that they are not correlated each other.

For an n-octave band's audio signal, the similarity score is finally defined as in Eq. 3.

$$S_{x,y} = \frac{1}{n}\sum_{i=1}^{n} c'_{x_i,y_i}(l). \quad (3)$$

When the frequency domain features are important as well as time-domain ones, standard classification techniques can be applied.

In the following, we briefly discuss Machine Learning Techniques used for classification, Statistical Classification Methods, and finally, the Robust Sound Hash (RSH) technique.

1) Machine Learning Techniques:

We briefly describe the most important machine learning techniques used in the scientific contributions included in this survey, i.e., Support Vector Machines (SVM), Random Forest, and k-Nearest Neighbors (kNN), as well as the widespread k-Fold Cross-Validation technique. For interested readers, we recall that a comprehensive study of the most useful machine learning techniques for audio analysis is available in [20].

Support Vector Machines (SVMs) [21] represent a supervised learning model for classifying data, i.e., assessing which pre-defined class is more similar compared to a new set of data. As a supervised learning technique, starting from a set of *training data* labeled on purpose as belonging to one or more categories, a SVM can classify new data as being part of a class or another. In addition, a set of unlabelled data, namely *testing data*, are provided for classification. Assuming s categories, the SVM internally builds a $s-1$-dimensional hyperplane in an s-dimensional space leveraging the labeled data, and maps the unlabelled one in this space. The output classification is provided as the category whose training points are the nearest to the testing data.

Random Forest [22] is an ensemble supervised machine learning technique, built as a combination of decision tree predictors. As an ensemble learning technique, the provided classification is the result of a decision taken collectively, from a large number of classifiers. The idea behind ensembles classification is based upon the premise that a set of classifiers can provide a more accurate and generalized classification than a single classifier, thus being less prone to over-fitting. When using Random Forest classifiers, each classifier is a tree, and each tree depends on the values of a random vector independently sampled, assuming the same distribution for all the trees in the forest.

K-Nearest Neighbors (KNN) [23] is a supervised machine learning algorithm able to solve both classification and regression problems. This methodology is based on the observation that instances with similar properties are generally found close together within a dataset. The classification method determines the label of an unclassified object, by observing the classes of its nearest neighbors. Thus, the algorithm determines the single most frequent class label in the set of labels of the k instances nearest to the unclassified object.

K-Fold Cross-Validation [24] is an accuracy estimation method that allows evaluating how the results of a model can be generalized to an independent dataset. The main objective of cross-validation methods is to estimate the generalization of a model, i.e., to understand its accuracy in the classification of data that it had never seen before, minimizing the effect of overfitting problems. The method consists in partitioning the dataset into subsets, some of them (the training set) used to perform the training of the model, and the remaining ones to be used for validation (the validation set) or testing (the testing set) purposes. Overall, it is worth noting that the SVM technique leads to better performances when dealing with multiple dimensions and continuous features, achieving its maximum prediction accuracy when the sample size is large. The Random Forest technique, instead, performs better when dealing with numerical and categorical features. Random Forest returns the probability of belonging to a class, while SVM calculates the distance to the boundary. Without loss of generality, we notice that no single techniques can outperform other algorithms over all datasets. Thus, the main issue when dealing with classification problems is to understand the conditions where a particular technique can significantly outperform other algorithms on a given application [25].

2) Other Statistical classification Metrics:

In the context of audio signal classification, the typical features used for optimal audio classification are the time and frequency correlation. While the time-domain correlation features the same logic previously described, the frequency domain correlation can be expressed via the Pearson's correlation coefficient. Defining $X(k)$ and $Y(k)$ the FFT of the two audio recordings $x(t)$ and $y(t)$, with $k = 1...m$, respectively, the Pearson's correlation coefficient is defined in Eq. 4, as:

$$s_f(X,Y) = \frac{\sum_{k=1}^{m}\left(X(k)-\bar{X}\right)\left(Y(k)-\bar{Y}\right)}{\sqrt{\sum_{k=1}^{m}\left(X(k)-\bar{X}\right)^2\sum_{k=1}^{m}\left(Y(k)-\bar{Y}\right)^2}}, \quad (4)$$

where $\bar{X}$ and $\bar{Y}$ are the mean of the vectors $X(k)$ and $Y(k)$, respectively [26].

Another coefficient, widely used to decide whether a prediction is correlated to the data or it is closer to a random guess, is the Matthews' Correlation Coefficient (MCC), defined as in Eq. 5 [27].

$$|\text{MCC}| = \sqrt{\frac{\chi^2}{n}}, \quad (5)$$

where χ is the chi-square [28] of the variables representing the two audio recordings.

Some other metrics can be used in descriptive statistics, such as the largest and the smallest observation values, defined as the values of the greatest and least elements of a sample, respectively.

A few classification systems used time-domain features in combination with other ones, since alone they are often not enough descriptive [29]. An example is the classification system presented by the authors in [30], which uses the duration of harmonic segments as a feature to identify human speech on audio samples.

Table II provides a summary of the most used correlation/similarity classification techniques in the scientific contributions included in this survey.

Classification Technique	Description	Time Domain	Freq. Domain
Pearson's Correlation Coefficient	Expresses the frequency domain correlation.	✗	✓
MCC	Check if prediction is correlated to the data.	✗	✓
Similarity score	Evaluates the similarity between two sounds.	✓	✗
MFCC	Represents the short-term power spectrum of a sound.	✗	✓
Peak Ratios	Considers the peak values of two variables.	✓	✓
Specific Band Energy	Considers the energy value in a specific octave band.	✗	✓

TABLE II
MOST POPULAR CORRELATION/SIMILARITY CLASSIFICATION TECHNIQUES.

3) Robust Sound Hash:
Another technique used to compare audio recordings is the Robust Sound Hash (RSH), especially when the exact matching between the audio recordings is desired. Ideally, to compare two audio recordings, a viable solution would be to compute the digest of each audio recording samples via a standard cryptographic hashing function, and to verify the matching of the two digests. However, as per the basic properties of cryptographic hashing functions, even small differences in source files would lead to totally different digests [31]. Thus, in the context of audio files, where the background random noise is always present, they are not applicable. A valuable alternative is the use of comparison mechanisms that change slightly in response to a minor variance of the input. An example is the Robust Sound Hash (RSH) technique, designed in [32]. The RSH divides the input file in different frames, each containing 1s of recording. Then, it applies a specific function to this digest, and it outputs a fixed-length string representing the sampling interval. The digests produced for each time frame are concatenated to form the final output. Then, the similarity between the two audio files is computed as the *Hamming Distance* between the two digests, i.e., the number of dissimilar bits. Because of the unique features of the function used in signal processing, similar audio recordings will have low dissimilar values. Thus, a threshold can be established to determine if two hashes are similar enough to be considered as matching recordings.

E. Masking Sounds

The *Sound Masking* technique is frequently used in audio-based defense systems to protect against unauthorized eavesdropping. It inherits its core logic from the *Friendly Jamming* technique, widely used in the wireless security research domain to hide sensitive information or block any other communication other than the authorized ones [33], [34]. The main logic of the sound masking strategy is illustrated in Fig. 5.

The audio leakage from a system, e.g. a keyboard or a vibration-based communication system, as thoroughly described in Section VI, can lead to the identification of privacy-sensitive information, such as the pressed key or the sound's

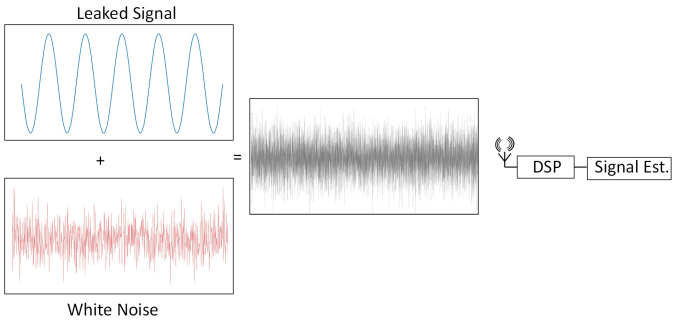

Fig. 5. Logic of sound masking mechanism. White noise produced on purpose sums up to the signal leaked from a given device, in order to corrupt it and make it unrecoverable to an adversarial Digital Signal Processing logic.

vibration. To mislead an external attacker, white noise is produced typically by an additional device coupled with the main communication channel, and it is emitted on purpose to taint the information signal. The malicious device, getting the mixed-signal, has the hard task of separating emitting components through Digital Signal Processing (DSP) techniques to estimate the useful signal.

F. Separating Sounds: The Cocktail Party Problem

The audio masking strategy described in Section II-E is effective only if an external third-party cannot perform a separating attack, i.e., it cannot decouple the jamming signal from the leaked information signal. Indeed, an adversary could perform such an attack in two different phases. In an online phase, it could gather the mixture signals via multiple microphones arranged randomly in the environment. Then, it could try to estimate the two addends, i.e., the information signal and the jamming signal. Assuming to have two different receivers and two different signal components s_1 and s_2, the mixtures x_1 and x_2 recovered by the microphones are expressed by Eq. 6.

$$x = H \cdot s + e \rightarrow \begin{bmatrix} x_1 & x_2 \end{bmatrix} = \begin{bmatrix} h_{1,1} & h_{1,2} \\ h_{2,1} & h_{2,2} \end{bmatrix} + \begin{bmatrix} e_1 & e_2 \end{bmatrix}, \quad (6)$$

where e_1 and e_2 represent the noise associated with each sampling of the audio signal by the receivers.

In the literature, this task is commonly referred to as the *Cocktail Party Problem*—in an analogy with the task of separating the linear mixture of different voices at a cocktail party, where voices of different tones overlap each other [35]. The class of approaches used to solve this problem is called Blind Source Separation (BSS) [36]. There are many BSS algorithms available in the literature, each one based on specific assumptions on the signals involved in the mixtures [37]. One of the most successful is the Independent Component Analysis (ICA) technique, whose goal is to express the mixed signal as a linear combination of non-Gaussian components, in a way that they are as much statistically independent as possible [38]. Regarding Eq. 6, this is equivalent to find an un-mixing matrix W, such as the following Eq. 7 is verified.

$$\hat{s} = \begin{bmatrix} \hat{s}_1 & \hat{s}_2 \end{bmatrix} = W \cdot x = \begin{bmatrix} w_{1,1} & w_{1,2} \\ w_{2,1} & w_{2,2} \end{bmatrix} \cdot \begin{bmatrix} x_1 & x_2 \end{bmatrix}, \quad (7)$$

Note that, if it is successful, ICA can recover the original mixing sources less than a scaling factor and a different naming (ordering) of the signals. However, in the context of audio mixtures, ICA may face applicability issues when the sources of the signals are very close to each other. In this case, the algorithm cannot precisely identify the exact number of the sources, or it confuses part of a signal with the other, resulting in degraded performance [39].

G. Distance Bounding via Audio

Like any other wireless signal, short-range audio signals can be used to provide location estimations. Specifically, looking at the signal emitted by the speakers and recorded back by the microphone, the emitting device could get a rough estimate of the device-user distance, further useful for advanced security tasks. Fig. 6 illustrates a sample scenario.

The simplest way to estimate the distance between a user

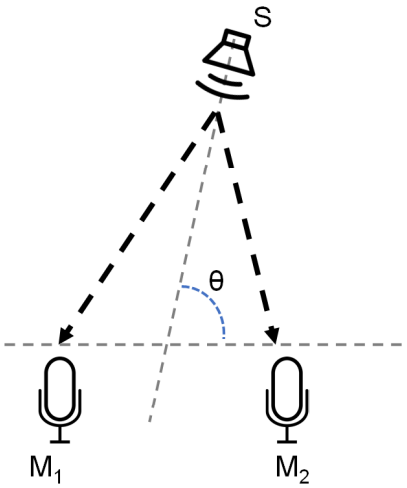

Fig. 6. Audio-based distance bounding. At least two receiving microphones M_1 and M_2 are used to differentiate the timestamps of the signals received from a single source speaker S, and to estimate the distance between the transmitter and the receivers.

and a device is to recur to the timestamps of the emitted and received signals. Let us assume that a single receiver is available, namely M_1, that t_0 is the time in which the sound is emitted and t_r is the reception time on the microphone M_1. The Time of Flight (ToF) can be obtained as $ToF = t_r - t_0$. Defining c the speed of sound, conventionally equal to 340 m/s, the distance between the user and the device can be easily obtained as in the following Eq. 8.

$$d = c \cdot \frac{ToF}{2}. \tag{8}$$

While being extremely simple to implement, the results obtained through the above formula, unfortunately, are affected by large errors, especially because of the large inaccuracies in timestamps gathered on mobile devices, such as smartphones and wearables, caused by various delays introduced in the processing chain [40], [41]. Thus, two techniques are mainly used. Where the timestamps of the signal are the only available data to be used and the emitting frequency cannot be changed, the Direction of Arrival (DoA) technique is a viable option. DoA resorts to the time of arrivals of the signals on at least two receivers (in the case of audio signals, microphones), to determine the location where the source is. Assume t_1 and t_2 to be the reception timestamps of a given signal emitted by a source S and received by two receivers, M_1 and M_2, respectively. Defining $\tau = t1 - t2$, the angle of arrival θ can be computed as in the following equation Eq. 9.

$$\theta = cos^{-1}\frac{c\tau}{d}. \tag{9}$$

Eq. 9 provides a single direction over which the source is located. The process can be repeated for every couple of considered receivers, providing an equal number of directions and thus a precise estimate of the specific point where the source lies. At the same time, inaccuracies due to environmental noise or hardware imperfections could be reduced.

An even more effective strategy to achieve source localization, namely Frequency-Modulated Carrier Waves (FMCW), is discussed in [42], with reference to radar systems, and it is based on frequency sweeps. The basic idea of this approach is to estimate the ToF indirectly, via differences in the frequency between transmitted and received signals. FMCW is based on a given number of rounds, being the final result the median of the results provided by single rounds. Assume that f_0 is the starting frequency and f_1 is the ending frequency, while T_s is the time needed by the transmitter to span the spectrum range from f_0 to f_1 in linear time at a pre-defined speed. At a given time, the transmitting device (i.e., the speaker) emits a sound on a frequency, that will be received by the receiving device on a given frequency. Defining Δ_f the frequency shift, i.e., the difference between the transmitting and receiving frequency, then the ToF, namely Δt, can be computed as in the following Eq. 10.

$$\Delta t = \frac{\Delta_f}{f_1 - f_0} \cdot T_s. \tag{10}$$

Note that, Eq. 10 does not contain the reception timestamps, but the transmitting and receiving frequency. Given that the transmitting frequency is chosen by the transmitter while the receiving frequency can be precisely estimated via standard FFT, the resulting accuracy can be highly improved. Once the ToF is obtained, the distance device-user can be obtained as in the following Eq. 11.

$$d = c \cdot \Delta t. \tag{11}$$

While the solutions discussed above will be further discussed throughout the paper, many other techniques to achieve source localization using acoustic signals are available. We refer the interested readers to the contributions already available in the literature, surveying methods for acoustic source localization [1], [43], [44].

III. Audio Channels for Two-Factor Authentication

In this section, we provide details on the use of audio channels as a second factor for the authentication of a user to a system or a device. The background on multi-factor authentication technique is provided in Section III-A, while the literature overview and the comparisons are provided in Section III-B.

A. Background

Two Factor Authentication (TFA) techniques are becoming increasingly pervasive and diffused in the last years, driven by ever increasing password leakage events[1] [45]. In addition to traditional password-based mechanisms, Two Factor Authentication (TFA) techniques adds a registered token in possession of the user to further enforce authentication and verify that the user credentials have not been leaked. Nowadays, TFA is successfully adopted in the context of online banking [46], enterprise access [47], and mobile social networks [48], resorting to dedicated private solutions or commercial services, such as Encap Security [49], Duo [50] and Google 2-step verification [51], to name a few.

The most widespread form of two-factor authentication is based on the use of electronic tokens, uniquely tied to the owner, such as smartphones. In the first phase, the user inserts its credentials. After successful verification of the credentials, the server in charge of the authentication generates a One Time Password (OTP) and delivers it to the user. The OTP is then used, in different forms, to provide a unique response to the server and further validate the possession of a secondary device. For instance, in case the electronic token is a smartphone, the OTP is delivered via SMS to the registered telephone number and the user is requested to insert it as it is received. The diagram in Fig. 7 highlights the described interactions.

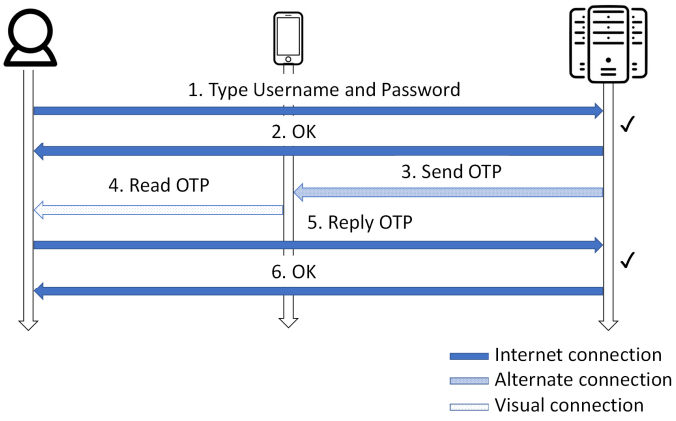

Fig. 7. Two-Factor Authentication: example of the interactions. In the first phase, the user performs the login using its credentials. Then, to further verify its identity, the server delivers a OTP to the mobile phone registered with the user, and it evaluates the reply provided by the user. If it matches the expected output, the physical individual behind the machine has both the credentials and the mobile phone registered by the user. Thus, it is authenticated.

[1]https://en.wikipedia.org/wiki/List_of_data_breaches

While the security offered by TFA techniques is undoubtedly higher than traditional mechanisms based on credentials only, its adoption, optional in most of the cases, is still not diffused. Indeed, most users still prefer the single authentication shot, mainly because of the extra effort required by legacy TFA techniques, always requiring explicit user interactions and not suitable for blind or visually impaired users [52], [53].
These drawbacks have motivated significant efforts in the last years, both by industries and academia, to develop more usable TFA schemes, eventually requiring zero interactions from the user. While some solutions based on the use of the Bluetooth [54] and Near Field Communication (NFC) [55] technologies exist, they require the involved devices to feature a shared alternative RF communication technology, not always available in modern devices [56].
In this context, the audio communication channel can provide an effective answer. Indeed, elements such as microphones and speakers are widely accepted, and already present in several devices, ranging from regular laptops to ultimate wearables.

B. Literature Overview

An early but successful approach to the use of audio for TFA was provided in [57], leveraging ambient sounds. The idea behind this approach is to demonstrate that the two involved parties, i.e., the client requesting authentication and the server in charge of authenticating requests, are in each other proximity, i.e., in the same physical environment. The resulting TFA mechanism, namely *Sound-Proof*, does not require any further user interaction, as the second factor for authentication is automatically enforced by triggering simultaneous registration of ambient sounds from the browser and the phone. The two-sides registrations of the ambient sounds are compared on the phone and, if significant similarities are found, the proximity of the client is established and the authentication is verified. The similarity between the two audio recordings/snippets is established via the *similarity score*, as outlined in Section II-D. Thus, the two parties access the audio channel only once, for quasi-simultaneous recordings. Then, the browser delivers its audio to the smartphone, via a regular data connection, for audio file comparison. This architectural choice is mainly applied due to privacy concerns. Indeed, if the evaluation of the audio recordings takes place on the server, a potentially malicious server could have access to the ambient sounds registered by the users, leaking sensitive information. Instead, Sound-Proof moves the comparison on the smartphone, while the server is only informed about the output of such operation. *Sound-Proof* only leverages ambient sounds, easily modifiable by a malicious third-party. Indeed, an adversary can deliberately make the phone create a given sound, or wait for the phone to emit a specific sound, such as a ring, a notification or an alarm sound. In these cases, the ambient audio fingerprint is indeed easily predictable, and it would allow an adversary to overcome physical barriers to provide a successful attack. This is the rationale supporting the Sound-Danger attack disclosed in [58], where the authors discuss several smart attacks via induced sounds on the phones, leading to a maximum of 83.2% efficacy in compromising user accounts leveraging the Sound-Proof technique.

An effective technique to overcome the Sound-Danger attack was provided in [59], with the assistance of a smart wearable device. In this scheme, namely *Listening-Watch*, the browser generates a random (i.e., unpredictable) audio signal, that is registered by the phone and the wearable device. The login succeeds if the audio recording on the wearable device contains a specific pattern and is similar enough to the recording of the browser, thus ensuring the proximity of the wearable to the browser. The similarity between the two recordings is established via the Similarity Score, as in [57]. The performance of *Listening-Watch* have been evaluated in an office environment, with different distances between the wearable and the browser, i.e., benign (30 cm), intimate (50 cm) and personal (from 50cm to 1 m), as well as with different levels of the audio signal, i.e., full volume (100% of the scale, 79 dBA), average volume (75% of the volume, 74 dBA) and low volume (50%, 67 dBA). A potential weakness of the *Listening-Watch* technique is to leverage audible sounds, emitted on purpose by the speakers. This could create potential usability issues, especially when the user needs to perform the login in specific areas, such as a library or during a talk. This issue is overcome by the authors in [60], employing ultrasounds ([18-22] kHz). The resulting technique, namely *SoundAuth*, does not disturb the surrounding audio environment. In addition, it employs the SVM technique for the audio comparison, resulting in a higher level of accuracy especially in attack situations. The security of *SoundAuth* has been evaluated directly against [57] in different environments, including an office, a desk with underlying music, a lecture and underlying TV audio noises. Thanks to the introduction of the SVM classifier, *SoundAuth* is characterized by lower levels of False Rejection Rate and False Acceptance Rate, leading to better performances.

Recently, the authors in [61] proposed *DoubleEcho*, a new context-based co-presence verification scheme that is effective in verifying that two (or more) devices are located in the same physical context. To this aim, *DoubleEcho* leverages acoustic Room Impulse Response (RIR), i.e., a time-domain representation of the modifications of a particular source signal at a given distance from the source in a particular physical location. Being dependant on both sound wave propagation properties and shapes and materials from the enclosure, the RIR is very difficult to be imitated or reproduced when not in the same physical environment of the participating devices. In this scheme, the devices initiate the recordings of the audio environment only when they are triggered by one of the participating entities, and they acquire audio samples for about 2 seconds. Given that the recording captures any sound from the surrounding environment within a half meter distance, the scheme could be not fully privacy-preserving: in fact, chances are that the recordings include background information, that could lead to the identification of the environment where the devices are, or to the leakage of additional private information. The scheme has been extensively evaluated, considering different smartphones and tablet brands, and a variety of public and private environments (including offices, kitchen, corridor, classroom, meeting rooms, to name a few). *DoubleEcho* significantly improves the state-of-the-art in the mitigation of context-based context manipulation attacks, requiring a MITM attacker to be less than 50 cm apart from the intercepted devices. For distances less than 50 cm, given that successful co-located MITM attacks are still possible, the protection is delegated to the authentication of the channel between the involved devices, that can be either Bluetooth, NFC, or WiFi.

Audible sounds are also used in the recent proposal by the authors in [62]. In this contribution, the authors proposed a TFA technique to strengthen the authentication of a wireless key-fob to a car, leveraging ambient sounds to overcome relay attacks. More in detail, at the reception of a wireless input from the remote key, the car produces random sounds in the audible frequency range, and both of the involved entities record the audio environment. The key delivers its recording to the car, where it is compared against the locally recorded sound context, using the *similarity score* metric. If the two recordings are similar enough, the key is successfully authenticated; otherwise, it is rejected. The method has been evaluated both in an office and in a real car-parking scenario, with a minimum Equal Error Rate (EER) of 0.0013. It is worth noting that, as for the proposal by the authors in [57], the reliance on audible sounds exposes the method to the Sound-Danger attack. In addition, when the adversary is co-located and can retrieve the locally stored secrets from the key-fob, the protocol is also vulnerable to Man In The Middle (MITM) attacks. Finally, being based on ambient sounds recordings, the protocol is also not privacy-preserving.

Despite the improvement in performance and usability, the previous approaches still do not provide an answer for a co-located MITM attack. Indeed, an adversary that is co-located with the end-device can stealthily relay messages between the two entities, bypassing the proximity check, as shown in [63] and [64]. A solution to this attack is provided by the *Proximity-Proof* scheme, described in [65]. It overcomes the co-located MITM attack through the extraction of the physical-layer fingerprints of the speaker and the microphone, thus authenticating the device involved in the communication. Thus, the TFA is approved not only if the audio environment similarity is verified, but also if the speaker and the microphone fingerprints are matched against the well-known profile stored at the server-side. The fingerprints of the microphone and the speaker are created by evaluating the effect of the transmitter and the receiver on transmission at a given frequency, spanning from 18.1 kHz to 20 kHz, as detailed in [66]. More iterations further refine the process, creating a more accurate fingerprint. The drawback of the fingerprint usually stems from the large number of audio samples required to establish the authenticity of the involved devices. Indeed, while the other approaches described above use only a single audio access for each involved device, in *Proximity-Proof* the phone speaker continuously delivers the virtual packet for $m \geq 1$ times, where m is a pre-defined parameter. The second-factor login is validated only if the login browser can establish the authenticity of the remote device with a probability $(1-p)^m$, p being the probability that the packet can be successfully decoded. As in [60], *Proximity-Proof* leverages inaudible sounds, coupling them with the Orthogonal Frequency Division Multiplexing (OFDM) technique and the ON-OFF modulation scheme,

while a simple Euclidean Distance method is used for audio comparison.

Tab. III wraps up the above discussion and summarizes the most important features of the described approaches. It is worth noting that the performance of the described approaches are reported using either the EER, or the percentage of False Positive and True Negative, or the Overall Error Rate (ER).

C. Lessons Learned

Our study, summarized by the comparison in Tab. III, highlights some important *lessons learned*, discussed in the following.

- **Reliance on Additional Communication Technologies.** First, we notice that a common feature of all the discussed approaches is to couple the audio channel with another (trusted) communication technology: this is used to transfer the audio recording to the device that is in charge of the comparison, and this step cannot be avoided or replaced. In some cases, such as in [61], the auxiliary communication technology is used also to protect against classical attacks against authentication protocols, such as MITM attacks.
- **Protection Against Relay Attacks.** The schemes discussed in Section III-B are based only on the use of the audio channel. Thus, the protection against MITM attacks, such as relay attacks, is typically delegated to the additional communication technology used to exchange data between the two devices. Alternatively, the actual literature also provides several Multi-Factor Authentication (MFA) schemes, combining the audio channel with additional sensing modes. As described by the authors in [67], pure audio-based schemes could be vulnerable against *relay attacks*, where an active adversary relays messages between two users, cheating them about the sharing of the same physical context. In this scenario, combining the audio channel with further sensing modes, such as WiFi, GPS, and Bluetooth could increase the co-presence assurance, and thus, the robustness of authentication schemes. An example in this direction is the work by the authors in [67], where the authors experimentally demonstrated that combining multiple sources improves the robustness of authentication schemes if compared with the use of the audio-only mode. By using the MCC statistical measure (with values between -1 and 1), the audio-only mode achieved values of 0.715, while fusing it with Bluetooth and GPS improves the performance up to values of MCC of 0.978. In the following contribution in [68], the same authors integrated their solution in a real-world application, namely *BlueProximity*, and evaluated the effectiveness of the resulting system against relay attacks.
- **Privacy Leakage.** We also remark that, while being effective and valuable approaches, *Listening-Watch*, *Sound-Auth*, and *Proximity-Proof* trade-off protection against active attacks with privacy. Indeed, moving the generation of the sound on the server enables it not only to assess the similarity of the audio recordings, but also to further use the collected audio recordings. Potential privacy leakages can emerge if the server is curious, i.e., it looks for sounds leading to the identification of the ambient where the user is located, extracting sensitive information. We also highlight that potential privacy leakages are increased when multiple sensing modalities are used.
- **Performance Issues.** Finally, as experimentally demonstrated by the authors in [69], the effectiveness of audio-based TFA schemes is strongly dependent on the particular physical context where they are applied. For instance, considering the *Sound-Proof* scheme proposed in [57], the authors in [69] discovered very different values of EERs not only in different physical contexts (e.g., car, static office, and office with mobile heterogeneous devices), but also when each physical context is in a different condition. For instance, when deployed in a car, *Sound-Proof* achieves an equal error rate of 0.071 in a city scenario, while its performance drops down to EER values of 0.124 when the car is parked, by using the same setup of the protocol parameters. Thus, the deployment of audio-based TFA schemes should be carefully evaluated by the system administrator based on the particular operational context, and its main parameters should be carefully tuned to maximize its performance.

IV. Audio Channels for Secure Device Pairing

In this section, we shed lights on the use of audio channels for secure device pairing between devices. The background on the device pairing technique is provided in Section IV-A, while the literature overview and the comparison between the discussed approaches are provided in Section IV-B.

A. Background

The research area widely known as *Secure Device Pairing* refers to the authentication of unfamiliar devices, i.e., the bootstrapping of secure communication between two wireless and possibly constrained devices, in a way to be robust against eavesdropping and MITM attacks [70]. Common pairing operations are, e.g., the association of a Bluetooth mouse, keyboard or headset with a laptop or another communication equipment, or the coupling of a smartwatch with a handheld device.

The best-known approach to solve such a problem is the standard Diffie-Hellman (DH) protocol. In the DH scheme, two entities sharing no prior secrets can establish a shared key resorting to exponentiation operations [71]. However, because of the lack of any shared knowledge, the authentication of involved parties is missing, and thus, MITM attacks are possible. This is the reason why several enhancements of the DH protocol have been proposed, leading to authenticated DH schemes [72].

However, in the context of *Secure Device Pairing*, traditional cryptography-based solutions are typically unsuitable. Indeed, the involved devices do not share anything: there is not any shared secret, no Certification Authority (CA) nor

Scheme	Feature	Audio Analysis Strategy	Sound-Danger attack robustness	Privacy pre-serv.	Number of Audio Accesses	Co-located MITM attacks robustness	Performance Evaluation		Perf.
							Env.	Dist.	
[57]	Ambient Sound	Correlation (Sim. Score)	✗	✓	1 per device	✗	Office, Music, Lec-ture, TV	Desk, Pocket	$EER = 0.07$
[59]	Audible Sound	Correlation (Sim. Score)	✓	✗	1 per device	✗	Office	Intimate, Personal, Benign	$EER < 0.01$
[60]	Ultrasound	SVM	✓	✗	1 per device	✗	Office, Music, Lec-ture, TV	Desk, Pocket	$EER = 0.14$
[61]	Audible Sound	Room Impulse Resp.	✓	✗	1 per device	✗	Office, Kitchen, Corri-dor, Class-room, Rooms, et al.	Room area	$FP, TN \in \{0.089 - 0.106\}$
[62]	Audible Sound	Correlation (Sim. Score)	✗	✗	1 per device	✗	Office, Parking	Personal, Benign	$EER = 0.001$
[65]	Ultrasound	Euclidean Distance	✓	✗	m per device	✓	Office	N/A	$ER = 0.02$ (at 50cm)

TABLE III
COMPARISON BETWEEN TFA APPROACHES BASED ON SHORT-RANGE AUDIO CHANNELS.

Trusted Third Party (TTP) to leverage for assisted operations. As suggested by the authors in [73], a theoretical solution perfectly suitable for this case would be the establishment of a common standard Public Key Infrastructure (PKI), trusted by all the devices. However, this would require all manufacturers to agree on a set of features, as well as endless revision processes, leading to large manufacturing delays. Thus, the human involvement in the process is necessary to confirm that the two devices are indeed communicating, and no other party is impersonating one of the legitimate entities. Given the cited premises, a challenging problem in this context refers to minimize the effort of humans in pairing operations.

Many solutions are available in the literature, already summarized by the authors in [74] and [72]. Common solutions are based on Out Of Band (OOB) channels, such as a camera or infrared links. However, these schemes could face deployment issues in some specific contexts, e.g., when the users are visually impaired, there are adversarial ambient light conditions, the area is security-sensitive (military areas) or, finally, the involved devices are not equipped with the necessary external modules, such as a camera or an infrared module.

In these scenarios, the usage of the audio channel could provide enhanced and peculiar features, emerging as a complimentary OOB channel to achieve a secure association between unfamiliar devices.

B. Literature Review

The idea of using the audio channel for secure device pairing was launched by the authors in [75], through the *Loud and Clear (L&C)* scheme. It uses audio to provide human-assisted device authentication, relying on the use of the spoken natural language. Specifically, *L&C* covers four possible use-cases, three of them are based on the emission of an audible and syntactically correct sequence by at least one of the devices. If the other entity involved in the pairing process has a speaker, too, it reproduces the detected word and the human has only to compare the two sequences, and to decide if they represent the same word. In case the other device does not have a speaker, the corresponding word is displayed as a text on a screen, and the user has to compare the text with the emitted sound. To reduce the number of audio channel accesses, *L&C* represents the authentication object as a syntactically correct text (usually non-sense), hashed to generate a fixed bit-string of H bits. Then, the hash is divided into 10-bit sections, and each section is mapped on a word. Thus, $H/10$ words are emitted, resulting in a completion time of about 32 seconds. Even if the scheme leverages an alternative communication channel between the two devices, no security assumption is made on this channel. Note also that the audio communication channel is not secret nor authenticated: thus, eavesdropping is always possible. The assumption about the underlying secrecy of the audio communication channel is instead at the basis of the proposal in [76], using a low-frequency audio channel to perform the pairing between an Implanted Medical

Device (IMD) and a tag reader sharing an unsecured Radio Frequency Identification (RFID) connection. Given the audio channel is assumed to be a secret as well as authenticated OOB (AS-OOB), the key is transferred directly via audio in approximately 2 seconds, leading to acoustic eavesdropping attacks if a microphone is placed very near to the human body.

Starting from the proposal by [75], a large number of audio-based approaches were proposed for devices pairing.

Another pairing protocol for constrained wireless devices was proposed in [77], with reference to devices featuring at least a screen and a microphone. The proposed approach, namely *Beep-Blink*, is based on the comparison of audio-visual patterns at the two sides of the communication to achieve secure device pairing. One of the two devices, featuring audio generation capabilities, encodes via audio a first Short Authenticated String (SAS) and an audio termination string, received by the second device. The other device acknowledges the reception of the two sounds through the blinking of a green led for the first string and a red led for the termination string. The user has just to compare the approximate matching of the two actions, i.e., beeping and blinking, to establish the association between the two devices. Indeed, the security of the scheme is based on the underlying security of the SAS strings.

A simple yet effective technique for secure device pairing on smartphones is provided in [41] and further refined in [78]. These contributions proposed the *Point& Connect* scheme, namely *P&C*, suitable for the pairing of smartphones. Specifically, it leverages a collaborative scheme based on a distance measure between involved devices based on acoustic signals, thus requiring only a speaker and a microphone. Indeed, in a scenario where multiple devices coexist, *P&C* allows for the selection of the target device through an intention-based mechanism where the initiator shakes and points his/her device towards the selected target destination device. The pairing intention is captured via an acoustic-based distance measuring technique, where the initiator emits two sequential chirp sound signals. At the receiver side, the intended receiver will be the device exhibiting the largest distance reduction, as it is on the Line of Sight (LoS) to the initiator's position. This can be detected via the Received Signal Strength Indicator (RSSI), pairwise ranging techniques or solutions based on the Time of Arrival (ToA). The proposed technique exhibits outstanding performance (error rate 0) when the positioning of the audio source is at an angle $\alpha = 0$ to the receiving device. Otherwise, the performance degrade quickly.

The audio channel is also employed in [79], as an OOB channel to assist pairing using an unsecured wireless communication channel. In the proposed approach, namely *UbiComp*, the short-range audio channel is used to convey information to verify cryptography elements delivered on the other, unsecured, wireless channel, essentially to guarantee the physical proximity of the end-device emitting the audio signal. A total number of 2 messages are delivered by the target device on the audio communication channel, consisting of a random number and a public key, and requiring 1.5s in total to be completely delivered, assuming a RSA key pair of 1024 bit. Of course, MITM attacks are possible, given that the public key of the devices is not authenticated. This is why the human is involved in the loop to verify that the sound is effectively emitted by the intended remote device.

Another solution leveraging the audio channel for pairing is the *HAPADEP* protocol [73]. It consists of two consecutive phases. In the transfer phase, triggered when the user pushes a button, one of the two devices transmits its cryptography materials (e.g, the public key) to the other one. Receiving this information, the remote device transforms the key in an audio sequence and plays it. The other device records and decodes the audio sequence, accessing the key. The audio transmission lasts for 3 seconds and, if the pairing is bidirectional, the scheme is repeated with inverted roles. During the following Verification Phase, the user is called to verify that the two sounds emitted by the involved devices are the same, and to press a key if a match is verified.

Extending the approach in [75], the same authors in [80] proposed a variant based on pure audio, avoiding the need for an auxiliary wireless channel. The basic idea leveraged by this technique is that if an attacker enforces a MITM attack by encoding a self-generated string, the legitimate user can detect the audio source and stop the execution of the protocol.

However, the contributions [81] and [4] confuted experimentally the idea that MITM attacks on the audio channel can be always detected because of the human perception of the communication channel.Hence, in [81], the authors demonstrated that systems based on self-jamming cannot fully reliably detect the presence of signals injected by a malicious adversary that is co-located with the victim, i.e., they are not able to face the so-called overshadowing attack. It is worth noting that facing the overshadowing attacks is more challenging than dealing with co-located MITM attacks described in Section III in the context of TFA techniques. Indeed, while solutions in the context of TFA can leverage shared secrets or certified fingerprints, this is not possible for the pairing scenario, where the devices are unknown to each other.

Further proofs of the effectiveness of the overshadowing attack are discussed in [4], where the authors also propose some countermeasures. The most effective is based on some security symbols, embedded within the transmitted data at a random location, such that the bit-string size of a generic security symbol corresponds to a single symbol in the bit-stream. For each symbol period, the samples above a given threshold represent security symbols, and they will be verified and discarded by the receiver. The remaining ones will be considered as data, to be demodulated and decoded. Thus, if an attacker transmits further data, it is very likely that the signal energy over the particular symbol will overcome the threshold and it will be discarded, causing a meaningless Denial of Service.

It is worth noting that, contrary to TFA schemes discussed in Section III, in the case of the audio-assisted pairing approaches discussed above the type of audio signal involved in the process is in the audible audio range. This is needed to allow the external user to confirm the correctness of the process. Contrariwise, some other approaches got rid of human involvement. As for the case of [78], a different strategy was proposed by the authors in [82], that aimed at eliminating

the human from the loop. It proposed *AdHocPairing*, an audio-based spontaneous device pairing application leveraging ambient sounds for secure key generation. The two devices synchronously record the ambient sounds for a duration of 6375 milliseconds, and they extract a fingerprint of such a record by taking its FFT in 33 sub-bands. To align the registration of the ambient sounds in two close but different locations, they map the recorded audio fingerprint on the code space of an error correction code, using the widest known principle of the Hamming Distance to identify the code-word that best matches the recorded audio fingerprint. One of the two devices, elected as the transmitter, use the best-matching key to encrypt a message, and sends it to the receiver. The receiver, having calculated the key employing the same principle, is then able to decrypt the information. In case it fails, it continues for up 10 attempts, taking every time the best key from the remaining ones having the shortest Hamming distance. While eliminating the human from the loop and resorting to the ambient audio-only, this scheme is more exposed against MITM attack, trading off security for usability. In addition, it is defined only for bi-directional pairing, while its specification in the context of unidirectional communications is not defined.

The authors in [83] provided a thorough study on how the audio channel state recorded by the devices can be used to generate a shared cryptographic secret between two devices in proximity, with minimal guessing probability by third-parties. Specifically, they proposed to extract audio fingerprints from synchronized recordings of mobile devices, to correct their hamming distance from specific pre-defined code-words using fuzzy cryptography schemes, and then to utilize these fingerprints as the seed for cryptographic keys generation among the involved devices. To provide a sufficient level of entropy to the key to be generated, this approach splits the full audio recordings, lasting 6.375 s, into 17 frames of identical length. Then, the FFT of each frame is computed and further divided into 33 frequency bands, where the respective energy values are computed. Overall, this operation results in a 512-bit fingerprint. Despite the effectiveness of the described approach, a 100 % match among fingerprints generated by different devices is very hard to happen. Indeed, the involved devices are characterized by different hardware, are separated by a small distance and they are not perfectly synchronized, thus introducing discrepancies in the fingerprint generation process. Thus, a fuzzy commitment scheme is implemented via Reed-Solomon codes, to generate a common secret among the involved devices. Using the combination of these strategies, the authors demonstrated how the ambient fingerprinting can be used for fingerprint-based authentication, even in the presence of an active attacker, able to disrupt the shared audio channel inserting audio messages on purpose. Specifically, the authors found that a minimum amount of 75% of correct bits in the reconstructed fingerprint are necessary to pair two communicating devices while, at the same time, providing sufficient security guarantees against eavesdroppers experiencing different audio contexts. A vast experimental campaign is provided to confirm the theoretical results, including office scenarios, situations in which attackers easily imitate the audio environments, and other busy environments.

Small variations of the approach in [83] have been described by the authors in [84] and [85]. The contribution in [84] focused on the practical issues related to hardware, environment, and time synchronization, necessary to achieve secret key generation via ambient audio fingerprints, and it provided several alternative options for features generation, as well as experimental results. The authors in [85], instead, exploited speech recognition techniques on free-form spoken interactions between the owner and the target device to identify precisely the devices to pair. Further, they restrict the pairing only to devices in proximity. The protocol was specified in two forms, based on the presence (or not) of a central authority, and it took up to a maximum of 4 seconds to complete, depending on the speed and length of the sentence used by the owner.

While previous approaches only rely on audio, another class of approaches extend the same concept over the whole *context* shared by co-present devices, coupling the audio channel with additional sensors readings.

The authors in [86] introduced the concept of *audio fingerprinting*, i.e., the unique ambient features characterizing a particular physical environment over time. Specifically, the approach proposed in this contribution combined audio-based key generation and a sensor-based approach, using the luminosity features of the surrounding environment. This concept is strengthened by the relationship with the time dimension: in fact, two devices can complete the pairing process only if the *ambient fingerprint* computed over the audio and the luminosity context is similar over a long time-span, i.e., multiple consecutive acquisitions. This innovative concept, namely *Sustained Co-Presence*, is integrated with a fuzzy commitment scheme, that can lead the two devices to estimate the same shared key only if, after the first initial pairing, they continue to share the same ambient context over the time. We notice that the scheme leverages an auxiliary secure communication channel to successfully and securely perform the fuzzy commitment scheme. In addition, acoustic eavesdropping is possible and effective to partially break the protocol, even if it should be coupled with the eavesdropping of the *luminosity* level in the underlying physical environment. The authors of this contribution report similarity values up to 91.8% for the audio channel, and of the 95.0% for the luminosity sensor readings, making the proposed solution an effective tool for the pairing of wearable and IoT devices.

The combination of ambient sound and light have been used also in [87], where the authors proposed a pairing scheme for wearable devices not equipped with external interfaces. Similarly to the proposal in [86], the scheme is rooted in the comparison between ambient sound and light features of the communicating devices, and it does not require strict time synchronization. As a distinguishing feature, the scheme proposed by [87] is triggered only when the user explicitly approves it, limiting the effectiveness of context guessing attacks. The protocol proposed in this contribution could achieve EERs of 0.1276 by using only ambient sounds, being however very susceptible to the underlying context features.

We recall that the actual literature provides a variety of context-based pairing protocols, possibly implicitly using also

the short-range audio channel as an element of the *context* experienced by pairing devices. To provide a reference example, the ambient sound is used by the authors in [88] as part of the features characterizing the physical environment where pairing is carried out. The protocol proposed in this contribution, namely *Perceptio*, does not specifically leverage the audio channel, as it allows two devices that do not share any common sensor to establish a shared key leveraging the physical context, removing any synchronization delay. Given that this survey is focused on short-range audio channels, and the research area of *context-based pairing* is just tangent to this contribution, we refer the interested reader to the surveys in [89] and [9] for an exhaustive list and comparison of *context-based* pairing approaches.

The techniques discussed above are summarized in Tab. IV, where differences in key characteristics of the approaches are highlighted

C. Secret Sharing via Short-Range Audio Links

When the pairing between two or more devices can leverage a secret, various heterogeneous techniques can be used. We hereby mention a few techniques using the audio channel to either share or allow the computation of a secret. Specifically, secret sharing is a well-known technique in the literature, used to distribute a secret between some entities in a network, such that subsets of these entities could decode and obtain the secret. For instance, if n is the total number of information pieces (namely, *shares*), and k is the minimum number of shares required to re-assemble the original information, the scheme is said to be a (k, n) threshold secret sharing scheme [90]. In this area, the authors in [91] further introduced Audio Secret Sharing Scheme (ASSS) techniques, leveraging short-range audible links to achieve secret sharing. These techniques were further studied by the authors in [92], proposing a new, secure but ideal audio secret sharing scheme. The main difference between a regular secret sharing scheme and an audio-based one as the one proposed by [92] lies in the numbness of human ears towards phase rotation of the audio signals. This property is used to design a scheme where no computation is required on the receiver side, i.e., on the human ear. The scheme is also demonstrated to be secure, in the sense that exactly k shares are required to reconstruct the signal. This work was further optimized by the authors in [93] and [94], concerning the security guarantees and the probability distribution used in the encryption of the original audio secret, respectively. In detail, it improves the result of previous works so that shares in ASSS schemes encrypting audio secrets should have bounded amplitude with probability 1. To achieve this objective, the authors use a normal distribution over a bounded domain, and evaluate the security of the resulting scheme. Finally, it is also worth mentioning the recent contribution in [95], further improving the state of the art by estimating the mutual information between secret and shares in a (n, n)-threshold ASSS.

D. Lessons Learned

The comparison summarized in Table IV highlights several outcomes and trade-offs, each discussed below.

- **AS-OOB Unsuitability.** Legacy approaches leveraging only the audio channel for secure device pairing, e.g., [76] and [79], were based on the AS-OOB assumption. However, the diffusion of audio-based pairing protocols and the large availability of computational resources onboard of modern devices rendered this assumption simply unfeasible. Thus, modern audio-based pairing protocols assume that, when used, the audio channel is accessible also to parties other than the legitimate ones.
- **Acoustic Eavesdropping Vulnerability.** As shown by the summary in Table IV, due to the open nature of the audio channel and the unsuitability of secret sharing, all the described approaches are vulnerable to acoustic eavesdropping attacks. Being passive and stealthy, the adversary could simply listen to the audio channel and, being in proximity of the involved devices, guess a large part of the bits of the secret used for device pairing. This vulnerability, along with the others described below, motivated the deployment of secure auxiliary channels between the pairing devices.
- **Overshadowing and Active Attacks Vulnerability.** As shown by the summary in Table IV, the overshadowing vulnerability discussed in [4] is an issue for all the approaches, independently from the human involvement in the pairing process. Especially in systems based on self-jamming, the overshadowing attack could be deployed by skilled adversaries, leading to the poisoning of the communication channel and a Denial of Service (DoS) on the pairing process. The authors in [4] also described some possible solutions to thwart overshadowing attacks, including physical-layer solutions based on Frequency Hopping Spread Spectrum (FHSS) and Direct Sequence Spread Spectrum (DSSS) techniques, and MAC-layer ones, based on the inclusion of security symbols within the legitimate audio signal. However, the deployability of these solutions within secure device pairing protocols is not straightforward, as all of them require the sharing of secrets between the involved devices (either for the synchronization of the frequency to be extracted among the possible ones, or for the positioning of security symbols within the audio data stream). The unsuitability of secrets sharing also creates the possibility of other attacks, such as relay and distance hijacking attacks. Relay attacks, being a kind of MITM attacks, assume an adversary that relays messages between the two involved parties, letting them believe to share the same audio context. Similarly, in distance hijacking attacks, the adversary jams the legitimate communication channel between the involved devices after the sharing of the information about the physical context, and completes the pairing process on behalf of one of the legitimate devices. These vulnerabilities, along with the acoustic eavesdropping previously highlighted, motivated the deployment of secure auxiliary channels between the pairing devices.

Scheme	Feature	Suitable for Unidirect. Pairing	Need of AS-OOB	Audio Channel Usage Time [s]	Acoustic Eavesdropp. Feasibility	Subject to Overshadowing	Auxiliary Secure Channels needed	Perf.
[73]	Audible Sound	✓	✗	3	✓	✓	✗	$ER < 0.05$
[75]	Audible Sound	✓	✗	32	✓	✓	✗	$ER = 0$ (within 10 cm)
[76]	Audible Sound	✓	✓	2	✓	✓	✗	N/A
[77]	Audible Sound	✗	✗	1	✓	✓	✗	$ER < 0.04$
[78]	Ultrasound	✓	✗	1.4	✓	✓	✗	$ER = 0$ (at $\alpha = 0$)
[79]	Audible Sound	✓	✓	1.5	✓	✓	✗	$ER = 0$ (within 30 cm)
[80]	Audible Sound	✓	✗	3.4	✓	✓	✓	$ER = 0$ (within 30 cm)
[82], [83], [84], [85]	Ambient Sound	✗	✗	6.575	✓	✓	✓	$ER <$ 0.001 (indoor env.)
[86]	Ambient Sound	✓	✗	$f \cdot w$ per device	✓	✓	✓	$EER =$ 0.12 (office env.)
[87]	Ambient Sound	✓	✗	6	✓	✓	✓	$EER =$ 0.1276

TABLE IV
COMPARISON BETWEEN PAIRING PROTOCOLS LEVERAGING AUDIO CHANNELS.

- **Secure Auxiliary Channels.** Pairing operations between unknown devices should not leverage any secret pre-shared between involved parties. At the same time, the security of the audio channel could be broken by the passive or the active attacks described above. Therefore, most of the available solutions leverage secure auxiliary communication channels to improve the overall robustness of the pairing scheme. Thus, in the first phase, the devices use the highest layers of the protocol stack and the auxiliary channel (Bluetooth, WiFi, and others) to establish a secure communication link. Then, the co-presence is established using the audio channel.
- **Hardware Heterogeneity.** In real deployments, the devices taking part in the pairing process could be very heterogeneous, being different not only in the number and the nature of embedded sensors, but also in the hardware chips providing sensing capabilities. These considerations are at the basis of several context-based pairing approaches, relying on physical properties that are invariant compared to the configuration of the pairing device. In this context, the microphones sensing the state of the audio channel could provide a different *intensity* and *timing* of a particular event. Thus, pairing approaches leveraging the audio channel should be able to remove time and intensity differences between pairing devices.
- **Usability.** Most of the analyzed approaches experimentally demonstrated that comparing the audio signals over an increased time improves the performance of the solution, reducing the overall error rate. However, increasing the audio channel usage time leads to an increase in the time needed for the pairing process, reducing the overall usability and transparency of the solution on the users' side. Thus, a trade-off is necessary between security and usability.
- **Additional Sensing Sources.** As also highlighted in the discussion related to TFA schemes, the actual trend is to leverage not only the audio channel for secure device pairing, but also additional sensing sources, such as light, GPS, Bluetooth, to name a few. The combination of all these data is leading to a switch from a pure audio-based secure device pairing to a *context-based* secure device pairing, where the audio is an element of the surrounding context shared by the devices. Many scientific contributions, such as [96], demonstrated the increased robustness of methods based on multiple sensing sources, through context-based proofs of presence that are hard to guess by adversaries.

  Despite this evolution, we believe that the study of pure audio-based schemes is still valuable, as the audio context stands as one of the most important elements defining the *physical context* shared by co-present devices at the time of the pairing. Thus, strengthening the co-presence de-

tection and increasing the robustness against well-known attacks on the audio channel (including eavesdropping and overshadowing) are still crucial requirements for any context-based approach leveraging the audio environment.

V. Device Authorization and User Verification via Short-Range Audio

In this section, we analyze the techniques available in the literature that use short-range audio to provide device authorization and user verification. Section V-A provides the background, while Section V-B reviews the current literature and provides final comparisons.

A. Background

Authorization is a fundamental security service that deals with the problem of identifying if an entity has the right to execute a given function on a given device [31]. A large variety of solutions exist in the literature, tailored for each system or network to be secured [97]. In this section, we focus on authorization mechanisms leveraging short-range audio links to grant or deny access to a system and physical devices.

A first area where audio is used to achieve authorization is in the user registration or authentication to websites, where it is crucial to distinguish a human being from automated software, namely a *bot*. This is a crucial task in many applications, including registration of users to websites, authorized responses from websites and command executions on actuators. This is a very common issue especially for website developers, where correct identification of humans and bots can achieve a valuable defense strategy [98]. Throughout this section, we refer to this issue as *user verification*.

Traditional mechanisms to distinguish between humans and bots were tailored for user registration to websites, and they are based on Completely Automated Public Turing tests to tell Computers and Humans Aparts (CAPTCHAs), widely used nowadays when registering a new account to a website or performing authentication to a known platform. CAPTCHAs identify humans and bots by challenging them with tests that are quite easy for humans, but hard for automatic devices. Examples include visual CAPTCHAs, based on images or videos, text-based or pointing-based ones, where the user has to click with its mouse in a specific location on the screen [99].

Another strategy is to recur to short-range audio challenges. Short-range audio is particularly effective for visually impaired individuals or people with motor limitations, as well as when the verifier is a speech recognition system, whose only interfaces with the external world are a microphone and a speaker.

B. Literature Review

In the context of user registration and authentication to websites, many different audio-based CAPTCHAs have been proposed, both based on human speech and ambient sounds, and a comprehensive overview of these strategies can be found in [12]. Without loss of generality, they assume a scenario in which the server is the *verifier*. It generates a complex sound, having features that are easy to be recognized by humans but hard to be identified by machines. The remote party, namely the *prover*, has to prove to be a human by providing an interpretation or decoding of the challenge. The overview in [12] tackles most of the solutions where the authentication/registration server is the verifier, and the user is the prover. However, many contributions in the last years have demonstrated the weaknesses of most audio-based CAPTCHA solutions [100]. Indeed, as reported in [101] the difference between human and computers audio capabilities is very small, much smaller than the difference between human and computer visual processing capabilities, and this makes it hard to design audio-based CAPTCHAs (further details will be provided in Section VI). The most successful recent techniques are provided by the authors in [102], where the authors proposed two different audio CAPTCHA schemes exploiting differences in auditory perception between humans and computers. The first technique explores the auditory perception of humans, especially when multiple simultaneous speakers occur. Indeed, humans defeat computers in isolating several similar acoustic streams from a mixture of signals, despite BSS solutions are even more closing the gap. To leverage this gap in the design of an improved audio CAPTCHA, multiple mixtures of speech signals are incorporated into an audio signal, by partially superimposing different words. To still maintain audio decodability features for humans, a variable time delay is introduced between two different words, and all speech pauses are filled with noise speech by multiple talkers. In the second approach, the authors exploit differences in humans' understanding of the natural language. The words in the CAPTCHA are now produced without any noise, and words are selected such as the probability confusion is minimal. To confuse attackers, artificially generated non-sense speech sounds are randomly inserted in the sequence, and the users are challenged to recognize meaningful words from semantically incorrect speech sequences. Note that both of the proposed techniques do not resort to any other communication channel rather than the audio means.

Also in the context of voice assistants, authorization is a crucial task. Many commercial voice assistants are available nowadays, including Alexa, Siri, Cortana, and Google Now, to name a few. Indeed, they are extremely convenient in scenarios where other kinds of interaction (visual-based or touch-based) require more effort or are unfeasible, such as in Driving Assistance Tools. Thus, ensuring that only authorized sources can trigger voice assistance tools is cumbersome. To provide a further authorization layer, the authors in [103] presented *VAuth*, a system guaranteeing that the voice assistant executes commands coming only from the voice of the owner. At the triggering of the voice assistant, the system collect accelerometer data from a wearable component mounted at the user's side, and it correlates them with signals coming from the microphone, via the MFCC. Only if there is a match with the profile of the user's voice, the command is issued.

Another application domain where audio channels are essential is in the context of audio-controlled Internet of Things (IoT) devices. In this dynamic ecosystem, voice interfaces are often provided to simplify the usage of IoT devices in the environment, e.g., by enabling voice-triggered lights

turn on, windows closing or opening, doors locking, and many other automated activities. Despite these devices are usually protected by ensuring the physical proximity of the triggering voice commands, often this is not enough. To protect audio assistants from unauthorized usage, the authors in [104] propose an authentication mechanism based on (at least) two microphones deployed randomly in the area. The main idea of the proposed system (*2MA*) is to locate the source of any audio source in the scenario, and to authorize the execution only of voice commands whose source is close to the user's mobile device. This is achieved by combining distance bounding, via standard Direction Of Arrival (DOA) mechanisms, and audio similarity, via a RSH function. In this case, another communication channel is used between the IoT device and the Mobile Device to exchange information. Thus, an initialization and key establishment phase are necessary.

Maintaining the focus on audio-controlled IoT devices, the authors in [19] provided a strategy to protect constrained devices from unauthorized actions triggering. Specifically, the authors tackled the issue of the triggering of IoT devices via other devices in the same area emitting sounds via electronic speakers. To distinguish between a human or an electronic speaker, they identify the presence of the sub-bass frequency excitation typical of any modern speaker. Aside, they also design a system filtering the noise of the environment, and computing experimental thresholds to decide if a human or a device is issuing voice commands. Thus, even assuming the attacker can manipulate audio commands, the electronic speaker will be always detected.

An overview of the approaches discussed above is provided in the following Tab. V.

C. Lessons Learned

The following take-away messages can be extracted from our discussion and comparison.

- **Semantic User Identification via Audio.** As highlighted by several studies, when taking the audio as the reference channel for authorization, differences between humans and bots are very small. When the usage of this channel is unavoidable (e.g., in the case of visually impaired users), the mental skills of humans are key to distinguish the nature of the user. Thus, users are asked to understand valid words within streams of characters, or to extract meaningful information from the audible sound. However, these operations have an impact on the overall usability of the solution, increasing the authorization time and overhead.
- **Authorization via Co-Presence Detection.** To protect audio assistants against unauthorized use, legacy techniques based on co-presence detection are used. Despite limiting the effective usage range of these devices, the integration of such physical-layer solutions helps to assess the effective presence of the entity issuing commands, avoiding relay and further MITM attacks.

VI. Attacks via Audio

In this section, we analyze several attacks performed through the audio channel. They are divided into active attacks, involving the delivering of audio signals on purpose, and passive attacks, performed by simply passively listening on the audio channel. In Section VI-A we describe the motivations behind these types of attacks, mainly related to exploiting Voice Controllable Systems (VCSs). We show that such attacks aim to achieve remote commands execution, as well as to violate user privacy exploiting the acoustic emissions of some devices. In Section VI-B we provide a thorough analysis of the most representative works that exploited the audio channel for malicious purposes, dividing them into the two aforementioned categories: active attacks (Section VI-B1) and passive attacks (Section VI-B2).

A. Background

The increasing diffusion of VCS, such as Voice Assistant (VA), is changing the way people live, offering several automation functions in both the consumer and the business realms. A VA is a software, usually installed on computers/smartphones, using different technologies including voice recognition, speech synthesis, and Natural Language Processing (NLP) to provide different smart services to the users. Thanks to a Voice User Interface (VUI), users can access the VAs and interact with them easily and quickly.

As shown in Fig. 8, the architecture of a VA includes (at least) three main components: voice capture, speech recognition, and command execution. The voice capture module records ambient and speech sounds, that are usually preprocessed and provided to the speech recognition system. This module can identify human voices containing possible commands, isolating them from ambient and noise sounds. Finally, the recognized commands are executed by the command execution module [105].

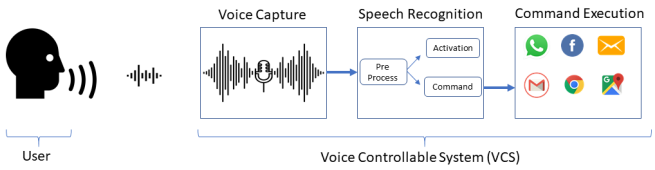

Fig. 8. Typical architecture of a VCS.

A VA usually works in two different stages: activation and recognition. The activation stage envisions that the VA continuously records ambient sounds, looking for a human voice. Indeed, the user activates the VA by pronouncing a predefined sentence (such us "Hey Google", or "Alexa"). If the predefined sentence is recognized, the VA enters the recognition mode. Other activation methods are possible, e.g., by pressing a physical button, or by opening a specific application. Once entered in recognition mode, the VCS converts voices into commands, thanks to NLP techniques.

The VCS technology is constantly evolving, and the VA market is progressing at an even higher pace. Indeed, the

Scheme	Verifier—Generating Sound	Prover—Interpreting Sound	Distinguishing Feature	Auxiliary Secure Channel Needed	Typical Scenario	Perf.
[102]	Server	Client	Partial overlapping of words	✗	User Registration on websites	Word Accuracy up to 90.51%
[102]	Server	Client	Natural Language Processing	✗	User Registration on websites	Word Accuracy up to 98.49%
[103]	Server	Client	Accelerometer Data	✓	Voice Assistants	Detection Accuracy up to 97%
[104]	Client	Server	Multiple Receivers	✓	Voice-controlled IoT devices	Detection Accuracy up to 97%
[19]	Client	Server	Sub-bass over excitation	✗	Voice-controlled IoT devices	Detection Accuracy up to 99.95%

TABLE V
COMPARISON BETWEEN AUTHORIZATION APPROACHES BASED ON SHORT-RANGE AUDIO.

widespread adoption of VA is a key factor behind the enormous growth of smart home devices and applications, with over 275 million VA devices expected to control smart homes by 2023, compared to the 25 million estimated in 2018 [106].

Following the rapid increase in popularity, VAs are gaining more and more attention from industry and academia, raising also important security concerns. Indeed, despite the undoubted usefulness and practicality offered by a voice interface, the lack of any inherent authentication mechanism is a cause of threats to the user's security and privacy, especially when using earlier versions of these softwares.

B. Literature Overview

The active attacks discussed in the recent scientific literature are mainly focused on attacking the VCSs. Indeed, these active attacks can be performed by sending audio signals on purpose to VCSs. These signals are neither unintelligible nor inaudible to humans, but still recognizable by voice capturing subsystems. Section VI-B1 provides a description of the most important contributions in this field, while Tab. VI summarizes and compares them. The most representative passive attacks discussed in the literature are described in Section VI-B2. Finally, Tab. VII highlights and cross-compares them across their main features.

1) Active Attacks: The majority of the attacks described below aims to cause the execution of unauthorized commands in a target device equipped with a VCS. These commands include malicious activities, such as data exfiltration (by sending of SMS or emails), money stealing (via payment-based SMS services), malware installation (by opening malicious web pages and starting downloads), and possibly others.

One of the first attacks against a voice assistant has been performed by the authors in [107], which proposed a bypassing attack exploiting the lack of any permission necessary for an Android application to access the phone speaker. The adversary model involves the use of a malware to invoke the voice assistant (Google Voice Search), bringing it to the foreground. Then, the malware plays a pre-prepared audio file containing specific commands, recognized and executed by the VCS. Indeed, many critical assumptions are necessary for the attack to be successfully executed. The victim's smartphone should be already infected with the malware, and the attack could be successfully completed only if the user is away from the phone. Indeed, if the victim is close to his smartphone, he can notice the ongoing attack by hearing the audio played by the malware through the speaker. Thus, he could manually block the execution of the command. Moreover, the attack has been tested only in a closed environment, without any external noise.

Another pioneer contribution that explored the feasibility of attacking a VCS using an audio signal unintelligible to humans was discussed in [108]. Authors provided a thorough study of the speech recognition mechanisms, investigating if differences in the ways humans and machines detect spoken speech could lead to exploitable vulnerabilities. They found that these differences can be easily exploited by an adversary to produce sounds that are recognized as a command by VCSs, while being perceived by humans as noise other than speech. This methodology increases the potential danger and the feasibility of these types of attacks. Indeed, these powerful attacks involve neither installing a malware nor performing any further interactions with the attacker and the victim's smartphone before the attack execution. Moreover, the audio signals injected by the attacker to activate and use the VA are stealthy, as they are unintelligible to human listeners. The attacker first produces an audio file containing a regular voice command that she wants to be executed on the targeted device. Then, the audio file is provided as input to specific software, namely an "audio mangler". This software removes all the audio signal features not used by the speech recognition module, but used by a human. The authors evaluated the real feasibility of this attack against the Google Now personal assistant, playing

the recorded commands using speakers located about 30cm from the targeted device. Their experiments involved several unauthorized actions, including activating the VA (Google Now), calling a number, sending SMS, and opening a website. Even if this work improved the previous attacks against VCS, some limitations are still present. These limitations involve the audibility of the unauthorized commands and the white box approach, which assumes some knowledge of the victim's speech recognition system by the attacker.

Some of these limitations have been overcome in [109]. In this paper, the authors proposed an attack against any VCSs, using a malicious audio signal coming from outside the victim's smartphone. The attack is realized with a few knowledge about the speech recognition system of the user, and it applies to a variety of scenarios. The attack is performed against the Google Now personal assistant software and the open-source CMU Sphinx speech recognition system, used by several VCSs. The authors significantly improved the state of the art by evaluating the feasibility of this attack under more realistic scenarios. In addition, they formalized a general method to produce machine-understandable speech that is, at the same time, rarely recognized by humans. The attack was performed in an isolated room, without background noise, playing the audio commands on a speaker located exactly 50cm from the microphone. This work practically demonstrated the feasibility to remotely attack a VCS, by activating a virtual assistant software and exploiting it to execute several commands without neither authorization nor privileges in the victim's system. On the opposite side, these attacks use hidden voice commands which are incomprehensible but still audible to humans. Thus, the user might notice the ongoing attack by hearing a noise.

To overcome this limit and increase the reach of the attack, a completely inaudible attack has been proposed by the authors in [105]. This objective is achieved by modulating voice commands on ultrasonic carriers ($f > 20kHz$), in a way to be inaudible to humans. The authors performed an in-depth study on the voice capturing subsystems (composed by microphones), investigating how they record audible sounds. They found that electric components such as amplifiers follow a non-linear behavior compared to the input signal features. Then, they found the same property on real microphone modules, observing that an attacker can exploit it to build an ad-hoc audio signal, containing the command. The voice is then modulated on an ultrasonic carrier right before the transmission. On the reception side, the sound is correctly demodulated on base-band by the receiver hardware. To demonstrate the feasibility of their methodology, the authors validated it using the most important speech recognition systems, such as Apple Siri, Google Now, and others. Specifically, they were successful in performing a variety of actions by simply injecting a sequence of inaudible voice commands, including accessing a malicious website, spying the victim accessing the image and the sound of its device, injecting false information, and many others.

Tab. VI summarizes our discussion and highlights similarity and differences of the approaches.

Among other scientific contributions in this category, it is worth mentioning the work in [110], where the authors provided an overview of the main technologies behind VCSs, including voice conversion and speaker verification. Furthermore, the most important spoofing attacks have been studied in a variety of scenarios, with a focus on the modification of a speaker's voice to emit sounds similar to the ones emitted by a different speaker, without modifying its content (namely, voice conversion spoofing attacks).

2) Passive Attacks: Acoustic side-channel attacks have been largely used to capture information from different devices, such as keyboards, printers, CPU, and even to identify information written by handwriting.

These attacks are mainly performed by placing a covert listening device in the physical proximity of the target device. This device simply records acoustic emanations from the victim device. Then, the recorded audio is processed to extract the desired features, further used to train a ML-based model. This attack has been proved to be feasible even recording the acoustic emanation through a Voice Over IP (VOIP) connection. The main goals include, but are not limited to, the victim's privacy violation, i.e., recovering the typed/printed text, or the violation of the intellectual property, i.e., recovering printed objects.

In the context of acoustic emanations analysis, the authors in [5] presented an attack against dot-matrix printers, intending to recover the printed English text. The attack includes recording the acoustic emissions with a microphone placed at about 10 cm from the targeted device. Indeed, at such a short distance, an attacker can recover up to 72% of printed words. Better performances, with a recovering percentage of up to 95%, can be obtained by assuming a contextual knowledge about the printing text. The attack can be divided into two steps. It involves a preliminary training phase, where the sound of printed words is recorded and used to train a machine learning-based model, and a second recognition phase, where printed English text is recognized. In the training phase, a word-based approach has been used instead of decoding individual letters, due to decay times and the induced blurring across adjacent letters. Interestingly, the authors found that most of the features characterizing printed sounds are located above the 20 kHz threshold. Thus, they identify the words in the recorded audio, analyzing the power spectral density above the 20 kHz acoustic threshold, and then spread the filter frequencies linearly over the whole bandwidth. Finally, they used digital filter banks to perform sub-band decomposition on each word. The authors tried also to perform the same attack against Ink-jet printers and laser printers, concluding that these printers technologies seem to be unaffected by this kind of attack. A similar attack against manufacturing systems has been investigated by the authors in [111]. The authors demonstrated that the sound emitted during the creation of an object effectively carries specific information about the process. Indeed, this information can be leveraged to reconstruct the printed item, even without any knowledge of the original design, violating intellectual property. The attack consists of

Scheme	Inaudible	Max attack range	Target Software	Attacker Capabilities	Attack Source	Tested Scenario	Performance
[105]	✓	165 cm	Apple Siri, Google Now, Samsung S Voice, Cortana, Amazon Alexa, Huawei Hi Voice	Commands Recognition, Activation	Remote	Office, Cafe, Street	Attack Accuracy from 30% to 100% according to the command
[107]	✗	Inside	Google Search App	Commands Recognition	Local	Noiseless Scenario	Attack Accuracy 100%
[108]	✗	30 cm	MFCC-based speech recognition systems	Commands Recognition, Activation	Remote	Quiet room	Being a feasibility study, the authors did not provide any performance indicator
[109]	✓	50 cm	Google Now, CMU Sphinx	Commands Recognition, Activation	Remote	Recorded Back-ground noise samples	Attack Accuracy over 90%

TABLE VI
COMPARISON BETWEEN ACTIVE ATTACKS VIA SHORT-RANGE AUDIO CHANNELS.

six phases. In the first *Acoustic Data Acquisition* phase, the acoustic emissions of the 3D printer are captured using a microphone placed 20 cm from the target. It is followed by a *Pre-processing* phase, involving the analysis of the audio to remove undesired noise. Then, a *Feature Extraction* phase is executed, where the time and frequency domain features commonly used in speech pattern recognition are selected to training the learning algorithm. Then, a *Regression Model* is applied to determine the speed of the printing. As per the definition, such a model consists of a collection of models each using a supervised learning algorithm for regression. Then, a *Classification Model for Axis Prediction* is used to determine the axis movement, and a *Direction Prediction Model* is applied to detect the direction of motion, using the frame energy of the audio signal. Finally, a *Model Recreation* process is applied to reconstruct the object using the output of the previous phases. The authors tested and validated their attack against a modern 3D printer, achieving an axis prediction accuracy of 92.54% and a length prediction error of 6.35% on a complicated object, such as a door key [111].

The same type of attack has been largely used also against keyboards. In this case, the attacker aims to recover the typed text, based on the observation that different keystrokes produce different sounds. The authors in [112] used neural networks with labeled training samples, identifying keystrokes using the FFT method and achieving an accuracy of 80%. This result has been improved a few years later by the authors in [113]. The authors revisited the attack methodology using unlabeled keystrokes samples and *Cepstrum* features, increasing the accuracy of up to 96%. A different approach is provided in [114], motivated by the observation that the sound emanated by a keystroke is related to its position on the keyboard. In fact, keys that are positioned close to each other emit more similar sounds if compared to those positioned far away. Moreover, the authors observed that working at the granularity of words instead of single keys, it is possible to leverage the statistical properties of the English language. Following these intuitions, the authors presented a dictionary attack able to recognize single words, having from 7 to 13 letters. Through specific signal processing tools, the authors achieved an accuracy of 73% over all the tested words. Furthermore, the attack does not require a training phase.

All these attacks are performed placing a microphone in the proximity of the targeted device. Despite this scenario poses a serious threat to the confidentiality of passwords and other sensitive text, the real applicability is limited. Indeed, the attacker should be located in the very near physical proximity of the user, being easily guessable.

To overcome this limitation, the authors in [115] presented a new scenario, considered the acquisition of acoustic information using VoIP protocols. The authors recorded keystroke sounds through a Skype call and demonstrated the feasibility of discovering what a user is typing, using a machine learning classifier based on MFCC features. They also investigated the impact that some variables have on the accuracy of their classifier, including the bandwidth and the presence of speech in the recorded audio.

The same attack is revisited by the authors in [6], with a focus on the recovery of random passwords and PINs. Instead of using a word-based approach as in previous work, the authors presented a methodology focused on single character detection. Thus, the accuracy of the attack depends only on a single character detection rate, avoiding the use of any language model and dictionaries. The training phase consists of the collection of single keystrokes (30 instances per key). Then, the authors used the trained model to detect randomly generated 6-character passwords, containing only lowercase letters (a, z), and 4 digits PINs. This is performed by using MFCC features extracted from password keystrokes, eavesdropped over the remote call. A similar methodology was used by the authors in [116], where the authors studied the vulnerability derived by audio emanations of keyboard typing. The authors showed that the keyboard eavesdropping attack performance is affected by a few variables, such as the typing style of the particular user, the data inserted by the user, and the adopted detection strategy.

Another type of acoustic side-channel attack was performed by the authors in [117] against PIN pad devices. This attack, called Differential Audio Analysis (DDA), analyzes the differential characteristics of the sound captured by two microphones placed inside the targeted device. Such a difference is then expressed as the cross-signal transfer function. Their experiments achieved the 100% of accuracy for certain devices, while only the 63% of accuracy was achieved when the target device does not produce the sufficient level of audible sound in correspondence to the press of a key.

Finally, a side-channel attack against handwriting has been investigated by the authors in [118]. The attack leverages the acoustic emissions of people handwriting recorded through a mobile phone. Indeed, this attack could lead to the leakage of personal information, eavesdropping, e.g, the sound derived from people filling out privacy-related forms. The authors presented a methodology based on audio signal processing and ML, able to recover the handwritten text. The audio is recorded with a smartphone placed in the same desk of the target user, about 30 cm far from the writer. Indeed, they achieved an accuracy of about 60% in word recognition, paving the way for further refinements and improvements. The approaches described above are summarized and cross-compared in Tab. VII.

C. Lessons Learned

The discussion on active and passive attacks carried out in the previous subsection allows us to highlight the following lessons learned:

- **Voice activation as a defense Tool.** All attacks against the VAs described in Section VI-B are based on the delivery of a malicious vocal command to the VCS of the target device. Then, the audio signal is interpreted as a legitimate command by the VA. However, to execute it, the attacker first has to activate the VA. The recent versions of common VAs are equipped with a *voice activation* feature, enabled by default, which allows the attacker to activate the VA with the same methodology, as the microphone is always active. Activating the voice assistant using a physical button, as in the oldest versions, would significantly mitigate the problem, requiring the attackers to use more complex attack mechanisms, e.g., activating the VA through a malware. These attacks are also promoted by the lack of defense techniques protecting VAs, such as using vocal authentication mechanisms, recognizing the voice of authorized users and enabling them only to use the VA.

- **Unintelligible/Inaudible audio commands.** As discussed in the previous section, attacks performed through inaudible commands could be very difficult to detect by the legitimate user. In fact, if the malicious signal is properly modulated, the audio command is perceived as noise by humans, while remaining intelligible by speech recognition systems. Moreover, if the attack is performed by using audio signals inaudible by humans, the attack becomes completely stealthy. This attack is enabled by the microphones commonly used on smartphones and other electronic devices, which are sensitive even at frequencies other than those audible to humans. Indeed, many microphones can sense sounds generated at frequencies higher than 20 kHz, even without any use case justifying it.

- **Extended attack range.** The first attacks against VA systems were effective only if performed near the target device. This consideration considerably reduces the real use cases of the attacks, making them difficult to execute. Unfortunately, as demonstrated by the authors in [105], by using the appropriate equipment and technologies, the attack range can be extended to larger distances, including 165cm. This wide distance, together with the inaudible features of the malicious audio command used for the attack, severely increases the severity of the threat. We highlight that, also in this case, the vulnerability is introduced mainly by the microphone, designed with properties higher than those required by common use cases. By decreasing the sensitivity of the microphone, it could be possible to limit the distance between the device and the audio source, without affecting the usability of the VA in regular use cases, usually requiring a distance of a few centimeters between the user and the device.

- **Audio as a side-channel.** The attacks described in Section VI-B2 demonstrate the feasibility of different kinds of privacy leakage, mainly due to the acoustic signals emitted by target devices. Although in many cases the performances are not very high, and usually some knowledge is needed about the victim's system, the feasibility of these attacks should not be underestimated. The lack of effective countermeasures highlights that further efforts by the research community are needed to mitigate this problem.

VII. Defense mechanisms using short-range audio

This section explores the use of audio channels as a defense mechanism against attacks previously described. Section VII-A provides the background, while SectionVII-B delves into the current scientific literature.

A. Background

The effectiveness of the attacks discussed in the previous section has motivated the design of ad-hoc defense techniques to limit their impact.

The idea followed by many authors consists of emitting sounds on purpose to protect the sound carrying information about a specific process. Without loss of generality, two approaches are used. The first is based on white noise emission, where an audio-enabled system is coupled with the target device and emit sounds that are flat in the frequency

Scheme	Eavesdropp. Position	Attack Vector	Target	Main Goal	Features	Algorithm	Performance
[6]	Same Room (15 cm)	Smartphone's Microphone	Keyboard	Password and PIN Detection	Sequential Minimal Optimization	MFCC	Attack accuracy 74.33%
[118]	Same Writing Surface (20-30 cm)	Smartphone's Microphone	Hand writing	Word Detection	SVM Based	FFT	Attack accuracy 55%
[5]	Same Room (10 cm)	Sennheiser MKH-8040 microphone	Dot-matrix printer	Recover Printed English Text	Undeclared	Sub-band decomposition	Attack accuracy 72%
[111]	Same Room (20 cm)	Condenser microphone Zoom H6	3D printers	Reconstruct Printed Object	Regression Model	Frame energy, Zero Crossing Rate, energy entropy, spectral entropy, spectral flux, MFCC	Attack accuracy 92.54%
[117]	Inside the Target Device	2 microphones	PIN pads	PIN detection	Yule-Walker Auto-regressive Method	Signal Transfer Function	Attack accuracy 100%

TABLE VII
COMPARISON BETWEEN PASSIVE ATTACKS VIA SHORT-RANGE AUDIO CHANNELS.

spectrum and overcome, in volume, the other ones. The second approach, instead, is based on the generation of sounds that are equal to the ones produced by the target device, thus poisoning the information received by the attacker.

The following discussion provides details about the usage of such techniques in specific sensitive conditions.

B. Literature Overview

A glaring example is provided by the authors in [119] and [120], in the context of vibration-based communications. These contributions tackle the PIN-Vibra method [121], used to transmit the keying material between two devices thanks to an ON-OFF vibration scheme. The receiver, equipped with MEMS motors, can detect the presence of vibrations and decode the key or, in general, a piece of information. While being exciting, these techniques have been demonstrated to be susceptible to eavesdropping attacks exploring the sound generated by vibrations [70], [122]. To overcome such attacks, [120] proposed Vibreaker, a system minimizing the leakage of information from the vibration-based communication channel via white noise masking. Specifically, Vibreaker couples the physical vibration with simultaneous audio emissions via the microphone of the smartphones involved in the communication. White noise is emitted by the microphone to cover sounds emitted by the vibrations, while low-frequency tones are further generated on-purpose and emitted to compensate for the partial inability of smartphone speakers to emit low-bandwidth sounds. On the attacker side, the best (minimum) Error Rate is 30%, that is very far to be acceptable for any valuable usage. Note that, while the user is not explicitly involved in the process (with undoubted usability gains), equipping the devices with a microphone causes a little overhead on the system, requiring not only motor accelerators but also audio-enabled devices. Similar properties were used by the authors in [123], in the context of medical devices. As before, the main communication channel is vibration-based and it involves a medical device and external equipment (smartphone or handheld). The proposed scheme, namely SecureVibe, is secured against acoustic eavesdropping derived from vibrations by generating appropriate masking sounds. As for the previous proposals, despite not requiring any active user involvement, this countermeasure forces a little overhead in the required components, as a speaker should be added (or coupled) with the medical device.The authors tested the effectiveness of their scheme against the ICA technique, and they verified that the scheme is secure up to a distance of 15 cm between the communicating entities. Instead, when the distance is higher than 15 cm, the ICA technique could be used to isolate the single signals, compromising the effectiveness of the technique.

Emitting white noise and fake keystrokes sounds are proposed by the authors in [124] and [6] to protect against attacks on keyboard typing. As thoroughly described in Section VI, the mechanical sound emitted by pressing and releasing keyboard keys can lead to their identification. To provide an effective defense, white noise and fake keystrokes are proposed, and their impact on the capability of the adversary to recover the correct key is evaluated. The best performance are achieved using fake keystrokes, where the attacker achieves a minimum error rate of 33%.

A similar defense mechanism is proposed by the authors in [125], concerning a zero-effort de-authentication system called ZEBRA, proposed by the authors in [126]. The system is made up of two components: a smartwatch and a software running on the PC, pre-paired wirelessly to each other. Both of them record the event they observe, including keyboard and mouse interactions, and the software on the PC compares continuously the recorded events, evaluating their matching.

When they no longer match, the user is assumed to be far from the PC and the user is de-authenticated. Several contributions, such as [127], demonstrated that ZEBRA can be cheated by an adversary that mimics the behavior of the user, e.g., via keystroke sounds emitted on-purpose. To provide a defense against such attacks, the authors propose to use the sound masking strategy, obtained by having the login terminals or a device placed in the surrounding environment to produce deliberate sounds, obfuscating the sound coming from the keyboards, and thus limiting the attacker possibility to reproduce the victim user's activities. Compared to the legacy ZEBRA scheme, the described defense mechanisms can improve the robustness to the aforementioned attacks up to the 70% error rate on the attacker side, complicating the effort of the adversary.

Defense-mechanisms using white-spectrum noise to thwart passive eavesdroppers are also used by [39] and [128]. The system proposed in these contributions consists of a sound-based communication channel, where speakers and microphones are used to deliver and receive information encoded in short-range audio messages. Given that the same speakers and microphones are used to emit the white-spectrum noise, this is the only technique where the defense mechanism can really be defined as *device-free*, as it does not require any additional device if compared to the unsecured scenario. While in the regular setting of the algorithm the ICA algorithm can isolate the waveforms in 90% of the cases at a minimum distance of 5cm, using random movements with a speed of 30cm/s the effectiveness of the ICA technique decreases to the 40%, complicating the task of identifying the correct waveform emitted by the legitimate devices.

A similar approach was used by the authors in [129]. Specifically, the authors proposed a system called Dhwani, i.e., an acoustics-based NFC system that uses the microphone and speakers on mobile phones within the same scope of legacy NFC communications, thus eliminating any specialized NFC hardware. As for [39] and [128], Dhwani uses the *JamSecure* technique, i.e., a self jamming strategy combined with self-interference cancellation at the receiver. This technique allows achieving secure communications from the information perspective among the communicating entities.

It is worth noting that, while the sum of audio signals could be a simple solution to avoid the leak of information towards trivial adversaries, smart and powerful adversaries can theoretically recur to algorithms such as ICA to decouple the summed signals, leading to the identification of the single components [38]. However, some experimental studies performed by the authors in [123], [39] and [105] experimentally demonstrated that if the two sound sources are very close to each other—e.g., within few centimeters—, the channel difference cannot be recognized by the receiving microphones, thus nullifying the efficacy of the ICA solution.

Another property of the audio signal that can be leveraged to provide enhanced defense tools is the proximity of two devices. Indeed, a well-known property of the audio signal is its quick attenuation over distance and physical barriers. Thus, the amplitude difference between the two recordings can be used to provide a rough estimation of the proximity of two devices. This principle is used in [130] to provide a solution for the Home Alone Wearable (HAW) attack. This attack is based on the assumption that users do not lock their wearable after their use and leave them unattended, hence enabling a malicious adversary to use the wearable to access information on the user device—e.g., a smartphone. To detect physical proximity between the user's smartphone and the wearable, the wearable triggers the phone to produce a sound, and both the devices record it via the microphone. The wearable then sends the recording to the smartphone via a Bluetooth connection, and the particular command issued on the wearable is executed only if the similarity score between the two recording exceeds a given threshold. It is worth noting that the user is not involved in the process, i.e., he/she does not have to state its explicit consent to authorize the information exchange. However, the wearable should be equipped with a microphone and the generated sound should be in the audible range, given that the microphones on the wearables usually cannot sample a tone in the ultrasound frequency range.

With the same objective in mind and using the same basic concepts, i.e., physical proximity, the iLock technique proposed by [131] recognizes the user's presence through the evaluation of the physical separation from the personal device, leveraging the changes in the wireless acoustic signals, and it immediately locks the hosting device if a threshold distance is overcome. iLock is based on the calculation of the ToF of the high-frequency acoustic signals (ultrasounds) emitted by the speakers. Being designed for smartphones, iLock does not require additional devices, given that microphones and speakers are already available in every commercial device, and it does not involve any further human effort. However, when applied in other devices, such as wearables, it could lead to an increase in the minimum features to be included in the system design. Considering the strongest attacker (i.e., an attacker that is closer to the phone than the legitimate user), iLock achieves a detection accuracy of the 60% when there is a difference of 0-270 degrees between the orientation of the user and the orientation of the attackers to the phone.

Physical proximity between communicating devices is also used by the authors in [132] to defend against Reader-and-Ghost attacks affecting NFC devices. In these attacks, a malicious reader tries to involve a legitimate NFC device in an unwanted transaction, piloted by a ghost adversary in another physical location, while the owner of the device wanted to perform a different transaction. To solve this issue, a transaction verification process is required, where the owner of the device specifically approves that transaction, and not a different one. To provide a defense mechanism against such a threat, the authors in [132] proposed a transaction verification strategy able to assure the proximity between a server and a mobile phone, leveraging the correlation between the related ambient audio recordings. The similarity between the two recordings is established via the computation of the cross-correlation, as described in Section II-D, and it does not require any user involvement. Without loss of generality, the work by the authors in [132] aims at detecting the co-presence between the involved devices. As discussed in previous sections III, IV, and IV-B, co-presence detection strategies, including context-

based strategies based on multiple sensing sources, are used for a variety of security-related tasks, including authentication, pairing, and access control. We refer the interested reader to the recent survey in [9] for a comprehensive overview of the strategies using sensing modalities other than the audio channel.

Tab. VIII summarizes the above contributions and highlights similarities and differences across the application scenarios.

C. Lessons Learned

The most important considerations and take-home messages arising from the above discussion are summarized below.

- **Protecting against Acoustic Eavesdropping.** In many cases, the audio channel can be used as a side-channel attack vector, to gain information exchanged using another channel, such as vibration. To protect against these events, dedicated sounds are introduced. These *synthetic* sounds aim to pollute the audio communication channel, decorrelating it with the main communication channel and thus nullifying its eavesdropping potential. The schemes used to achieve this objective can be manifold, e.g., based on white noise, random sounds, or dedicated sounds mimicking the ones produced by the main communication channel, such as fake vibrations or keystrokes.
- **Masking Sound Challenges.** The robustness of any audio-based defense scheme leveraging dedicated sounds emitted on purpose should be evaluated against the Independent Component Analysis (ICA) attack. Among the wide possible number of solutions to decouple signals emitted by different sources, ICA has been demonstrated to be the most effective one, and to be particularly successful when the emitting sources are sufficiently far each from the others. Therefore, masking sounds are suggested as an audio-based defense scheme in situations where the emitting sources are very close to each other, such as in the context of Implanted Medical Devices or voice-controlled IoT devices. When a consistent separation exists between the sources of the audio signal, the designer of the solution should ensure that the attacker could not deploy a sufficient number of receivers to be able to perform subsequent analysis.
- **Usability and Transparency of Audio-based Defense Schemes.** Any audio-based defense solution should deal with the resulting usability of the system. In the literature, the usability of the solution is often evaluated by looking at the requirement of additional components to the system, as well as to the explicit involvement of the user in the application of the defense strategy. Therefore, any valuable defense scheme should require neither any additional component nor the explicit involvement of the legitimate users. At the same time, possible distraction factors on the user's side should be reduced. For instance, when dealing with fake keystroke injection, despite the proposed defense scheme could be valuable and effective, the resulting distraction factors on the user could make the defense solution hardly usable.

VIII. Research Challenges and Future Research Directions

In the following, we identify the relevant research challenges related to each of the application domain analyzed in previous sections. We believe that the following discussion could be relevant for researchers actively working in the area, for inspiring new solutions and enhancing existing ones.

Privacy-preserving Two-Factor Authentication. The existing approaches, thoroughly described in Section III, have been demonstrated to be able to reach a very high level of usability, similar if not superior to existing token-based approaches. However, privacy concerns still represent a pressing issue. Indeed, leveraging pure ambient sounds exposes the whole scheme to biasing by malicious adversaries, as well as to co-located MITM attacks. A successful solution to these issues leverages the generation of random sounds on the server, but it forces the authenticating device to deliver the local recordings to the server itself, possibly leaking information about the surrounding environment. Thus, privacy-preserving TFA mechanisms based on short-range audio signals are still missing. Possible solutions to face these privacy issues could leverage innovative encryption techniques, such as Homomorphic Encryption (HE) strategies, widely emerging in the last years [133]. Indeed, the HE paradigm is a particular type of encryption, that can address the above privacy concerns by allowing a third actor to use encrypted data without the need to decrypt them.

In the context of audio-based TFA, the smartphone in possession of the user and the browser can first record the ambient sound, then encrypt the data recording using one of the many available HE techniques, and finally, deliver it to the server in the form of encrypted data. At the reception of the data, the server could compare the sounds recorded locally and remotely, operating over encrypted data, minimizing privacy concerns. These techniques have been already applied in various application contexts to similar privacy issues, such as in [134] and [135].

We remark that, despite FHE strategies are normally associated to heavy computational requirements, few contributions in the literature already demonstrated that the main source of the computational overhead is not related to encryption operations, but only to operations over the encrypted data [136]. In the short-range audio TFA scenario, comparisons operations over encrypted data are executed on powerful servers, that can reduce the time needed to provide a decision by allocating more resources for this task.

Thus, the browser and the smartphone could deliver encrypted data to the server, without any risk for potential privacy issues.

Despite the promising directions, however, the first Fully Homomorphic Encryption (FHE) scheme has been revealed only in 2009, and still needs further improvements and refinements to be practical on nowadays computing

Scheme	Audio Signal	Main Channel	Physical Property	Device free	Explicit User Involvement	Scope	Perf.
[119], [120]	White Noise	Vibration	Masking Sounds	✗	✗	Eavesdropping Protection	Attacker Best Error Rate 30%
[123]	White Noise	Vibration	Masking Sounds	✗	✓	Eavesdropping Protection	50% Error Rate up to 10 cm
[124], [6]	White Noise, Fake Keystrokes	Acoustic (Keystroke)	New sounds generation	✗	✗	Eavesdropping Protection	Keystroke Detection Error Rate 33%
[125]	White Noise, Music Sounds	Acoustic (Keystroke)	Masking Sounds	✗	✗	Eavesdropping Protection	Attacker Error Rate 70%
[39], [128], [129]	White Noise	Acoustic	Masking Sound (Eraseable)	✓	✗	Eavesdropping Protection	Eavesdropping Success Rate 40%
[130]	Fake Audible and Ambient Sound	RF	Proximity Detection	✗	✗	Anti-Theft	Maximum Attacker False Acceptance Rate 5.3%
[131]	Ultrasound	Physical Contact	Distance Bounding	✗	✗	Anti-Theft	Attack Detection accuracy 60%
[132]	Ambient Sound	NFC	Proximity Detection	✗	✗	Transaction Verification	Attack Detection Accuracy 100%

TABLE VIII
COMPARISON BETWEEN DEFENSE APPROACHES LEVERAGING SHORT-RANGE AUDIO CHANNELS.

platforms [137].

Lightweight Solutions to MITM Attacks. Despite the recent contribution in [65] provided a viable solution to overcome co-located MITM attacks, further work is still necessary. Indeed, running a fingerprint evaluation process as the one provided by the authors in [66] could not be viable on more constrained devices, both because of the high processing capabilities required to provide a classification in a reasonable time, and because of the high number of audio messages to be exchanged. Alternative solutions could leverage an enhanced selection of features to fingerprint the audio environment in a specific location around a device, further minimizing the area around the target that could lead to a co-located MITM attack. Here, the challenge also lies in the minimal complexity of the fingerprinting process on devices such as commercial smartphones, not featuring the same processing and storage capabilities of server platforms.

Non Invasive Audio-Based Pairing Schemes. In the context of pairing, reducing the involvement of the human user is the most pressing issue. Indeed, removing the human from the process could also pave the way to less invasive solutions, not requiring audible sounds. The resulting stealthiness of the paring process could theoretically broaden the application scenarios where audio-based pairing schemes could be performed. This shift, however, cannot prescind from auxiliary means to verify that the pairing process is ongoing exactly between the intended devices.

Facing Overshadowing Attacks in Constrained Devices. Novel solutions to face inaudible overshadowing attacks are highly required. Despite the solution based on security signals proposed in [4] is effective, it could be very energy and time demanding in specific contexts, i.e., involving Bluetooth and other battery-powered devices, where turning on the microphone or the speaker many times could be undesirable. In addition, requiring the sharing of secrets between the devices, such a solution is not practical when pairing devices previously unknown to each other.

Acoustic Eavesdropping Risks Analysis. Acoustic eavesdropping in the context of pairing deserves more attention. Indeed, even if the acoustic eavesdropping is a well-known threat, actually it does not seem a crucial one, given that the data delivered on the audio channel cannot be used to launch an attack. Indeed, it is worth noting that encryption techniques, as well as other security techniques, are very rarely applied to the audio signals, and their misuse could easily lead to severe threats. For instance, it is not

clear if simple *Replay Attacks*, performed by recording and playing back again the recorded audio signals, could lead to more severe threats when launched on audio-based pairing protocols.

Overcoming the Hardware Heterogeneity. The heterogeneity of microphones and speakers, as well as their different *tolerance* to the noise in different application scenarios, take part in degrading the performance of audio-based pairing schemes when moved away from a suitable application scenario. For instance, as shown by the authors in [69], the same Zero-Interaction audio-based solution report very different performance in indoor and outdoor scenarios. To overcome these limitations, researchers should design coefficient and comparison methods able to both remove inherent inaccuracies in the devices' hardware, and evaluate similarities and differences between recorded sounds in heterogeneous application domains.
In real deployments, the devices taking part in the pairing process could be very heterogeneous, being different not only in the number and the nature of embedded sensors, but also in the hardware chips providing sensing capabilities. These considerations are at the basis of several context-based pairing approaches, relying on physical properties that are invariant compared to the configuration of the pairing device. In this context, the microphones sensing the state of the audio channel could provide a different *intensity* and *timing* of a particular event. Thus, pairing approaches leveraging the audio channel should be able to remove time and intensity differences between pairing devices.

Securing Voice Assistants. Despite the relevant advances of the last years, securing voice assistants is still an issue. While actual solutions for authorizing the usage of voice assistants are referred to as *Authentication* techniques, in practice they only provide a probabilistic assurance that who is issuing commands is effectively authorized. Limitations still exist, e.g., the approach in [104] has an uncertainty area of ± 15 degrees around the target. Thus, it is very effective in the proximity of the source, but it loses efficiency when the source is far. Similarly, the speaker fingerprinting technique applied in [19] may become inaccurate when in the background there is not silence, but other sounds. This is a typical situation in navigation assistants or voice assistants running on smartphones. At the same time, biometric recognition techniques to identify the voice of a (set of) legitimate owner(s) have still received little attention from the community, and needs to be further investigated to prove its effectiveness. This is especially true when they are applied on constrained devices, such as IoT audio-controlled devices. Indeed, efficient and effective techniques are needed, possibly coupling already available solutions in an engineered fashion.

Sound Masking Effectiveness Evaluation. Currently, many audio-based defense techniques are based on *sound masking* strategies, described in Section II-E. However, some contributions such as [39] already provided evidence that this technique can fail to provide the desired masking properties if the sources of the audio signal, i.e., the genuine sound and the masking sound, are located sufficiently far from each other. This limitation seems to be a big issue in specific contexts. For instance, when protecting sounds emitted by keys on a keyboard, the speaker of the laptop and the keyboard could be sufficiently far from each other to allow an attacker to identify the location of the device emitting the particular sound. At the same time, the effectiveness of injecting fake sounds is still not widely assessed in practical conditions. Indeed, the effectiveness of these techniques should be carefully evaluated in the presence of attackers able to recognize and to tell apart sounds emitted by a speaker and mechanical sounds, recalling the technique used in [66].

Proximity Detection Schemes Improvement. Another defense tool leveraging audio signals is proximity detection. This approach refers to the similarity of ambient audio to assess co-presence between communicating devices. As shown in a recent contribution [138], it is feasible for a powerful attacker to manipulate the context and bypass the proximity detection evaluation by biasing the environment, e.g., inserting sounds or predicting the environment. Thus, audio-based proximity detection could be probably more useful if used in combination with other co-presence detection methods, in a way to force the attackers to realize more expensive and powerful attacks. Indeed, proximity detection could not provide a definitive assurance that the attacker cannot overcome the defense strategies. Thus, contributions are still needed in this area.

Attacks and Defenses via Directional Microphones. The large majority of attacks available in the literature and discussed in the previous section focused on omnidirectional microphones. Indeed, as per their definition, they can pick up sounds from almost every direction with the same performances. While this can be a useful feature in some scenarios, there are situations where an enhanced sensitivity towards particular sounds is desired. In this context, directional microphones can provide a meaningful solution. Indeed, a directional microphone is more sensitive to picking up sounds in certain directions rather than others. At the same time, specific types of directional microphones increase the reception range along a specific direction, as they can detect sounds emitted at a higher distance than regular omnidirectional ones [139], [140]. Indeed, these interesting features make them an attractive choice both for attacks and defense. On the one hand, attackers could increase the minimum required distance to launch an attack against a target device. On the other hand, directional microphones can implicitly reject attackers located outside the main reception lobe. Despite their evident advantages, the use of directional microphones for both attacks and defense is still not widespread yet. Hence, it represents a promising and interesting research direction to improve state-of-the-art results and to come up with innovative security tools.

Advanced Security Schemes via 3D Audio. Thanks to the very high pace of technological innovations, the first

3D Audio systems are hitting the market, with the promise to significantly enhance the sound experience of moving users and devices [141]. 3D audio recordings are created by placing a large number of microphones in the area. In this way, the sound of the scene is recorded at the same time from different positions, allowing to re-create the real sound experience of a moving entity. From the security perspective, 3D audio schemes could have a disrupting potential, both from the attack and from the defense side. To date, there are no studies in the literature that investigate this venue, that is probably the most novel and promising one in the short-range audio-based security context.

IX. Conclusions

In this paper, we have provided a thorough survey of mechanisms, applications, use-cases, and research challenges for short-range audio channels security. We showed that, thanks to enhanced usability features and low deployment costs, techniques based on short-range audio channels can be used as a means to achieve innovative and effective security services, such as Two-Factor Authentication, pairing, and device authorization schemes, to name a few. These properties are enforced by leveraging specific physical-layer features of the audio channels, such as distance bounding and physical proximity detection of devices sharing the same audio context. We have also shown that, if not integrated correctly, such methodologies could be subject to a variety of attacks. Thus, research is needed to improve defense solutions based on short-range audio signals.

Finally, we have also highlighted upcoming research challenges. The exposed challenges show that the development of audio-based solutions is still an exciting research area, and that it can be inspiring for researchers, industry, and start-ups, striving for innovative, non-invasive, and computationally lightweight means to enforce systems security.

Acknowledgements

The authors would like to thank the anonymous reviewers for their comments and suggestions, that helped improving the quality of the manuscript.
This publication was partially supported by awards NPRP11S-0109-180242, UREP23-065-1-014, and NPRP X-063-1- 014 from the QNRF-Qatar National Research Fund, a member of The Qatar Foundation. The information and views set out in this publication are those of the authors and do not necessarily reflect the official opinion of the QNRF.

References

[1] J. A. Belloch, J. M. Badía, F. D. Igual, and M. Cobos, "Practical Considerations for Acoustic Source Localization in the IoT Era: Platforms, Energy Efficiency and Performance," *IEEE Internet of Things Journal*, pp. 1–1, 2019.

[2] Z. Pan, Y. Ge, Y. C. Zhou, J. C. Huang, Y. L. Zheng, N. Zhang, X. X. Liang, P. Gao, G. Q. Zhang, Q. Wang, and S. Shi, "Cognitive Acoustic Analytics Service for Internet of Things," in *IEEE Int. Conf. on Cognitive Computing*, June 2017, pp. 96–103.

[3] MarketsAndMarkets, "Wireless Audio Market by Product, Technology (Bluetooth, Wi-Fi, Airplay, RF), Application (Home Audio, Consumer, Commercial, Automotive), and Region - Global Forecast to 2023," Tech. Rep., Jun. 2017.

[4] Q. Hu, Y. Liu, A. Yang, and G. Hancke, "Preventing Overshadowing Attacks in Self-Jamming Audio Channels," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–1, 2018.

[5] M. Backes, M. Dürmuth, S. Gerling, M. Pinkal, and C. Sporleder, "Acoustic Side-channel Attacks on Printers," in *Proceedings of the 19th USENIX Conference on Security*, ser. USENIX Security'10, 2010, pp. 20–20.

[6] S. Anand and N. Saxena, "Keyboard Emanations in Remote Voice Calls: Password Leakage and Noise(Less) Masking Defenses," in *Proc. of ACM Conf. on Data and Application Security and Privacy*, ser. CODASPY '18, 2018, pp. 103–110.

[7] G. Petracca, Y. Sun, T. Jaeger, and A. Atamli, "AuDroid: Preventing Attacks on Audio Channels in Mobile Devices," in *Proceedings of the 31st Annual Computer Security Applications Conference*, ser. ACSAC 2015, 2015, pp. 181–190.

[8] Q. Hu, J. Zhang, A. Mitrokotsa, and G. Hancke, "Tangible security: Survey of methods supporting secure ad-hoc connects of edge devices with physical context," *Computers & Security*, vol. 78, pp. 281–300, 2018.

[9] M. Conti and C. Lal, "A survey on context-based co-presence detection techniques," *arXiv preprint arXiv:1808.03320*, 2018.

[10] L. Nguyen and A. W. Roscoe, "Authentication protocols based on low-bandwidth unspoofable channels: a comparative survey," *Journal of Computer Security*, vol. 19, no. 1, pp. 139–201, 2011.

[11] G. Deepa, G. SriTeja, and S. Venkateswarlu, "An overview of Acoustic Side-Channel Attack," *International Journal of Computer Science & Communication Networks*, vol. 3, no. 1, pp. 15–20, 2013.

[12] Kulkarni, S. and Fadewar, H., "Audio CAPTCHA Techniques: A Review," in *Proceedings of the Second International Conference on Computational Intelligence and Informatics*. Singapore: Springer Singapore, 2018, pp. 359–368.

[13] P. Jayaram, H. Ranganatha, and H. Anupama, "Information hiding using audio steganography–a survey," *The International Journal of Multimedia & Its Applications (IJMA) Vol*, vol. 3, pp. 86–96, 2011.

[14] B. Gold, N. Morgan, and D. Ellis, *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons, 2011.

[15] I. R. Titze and D. W. Martin, *Principles of voice production*. ASA, 1998.

[16] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," *Journal of Supercomputing*, vol. 2, no. 3, pp. 138–143, 2010.

[17] M. R. Hasan, M. Jamil, and M. Rahman, "Speaker identification using mel frequency cepstral coefficients," *variations*, vol. 1, no. 4, 2004.

[18] N. Sato and Y. Obuchi, "Emotion recognition using mel-frequency cepstral coefficients," *Information and Media Technologies*, vol. 2, no. 3, pp. 835–848, 2007.

[19] L. Blue, L. Vargas, and P. Traynor, "Hello, Is It Me You'Re Looking For?: Differentiating Between Human and Electronic Speakers for Voice Interface Security," in *Proc. of ACM WiSec '18*, 2018, pp. 123–133.

[20] F. Camastra and A. Vinciarelli, *Machine Learning for Audio, Image and Video Analysis: Theory and Applications*. Springer, 2015.

[21] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[22] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[23] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine learning*, vol. 6, no. 1, pp. 37–66, 1991.

[24] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2. Montreal, Canada, 1995, pp. 1137–1145.

[25] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised Machine Learning: A Review of Classification Techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, pp. 3–24, 2007.

[26] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Pearson Correlation Coefficient*. Springer Berlin Heidelberg, 2009, pp. 1–4.

[27] B. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA) - Protein Structure*, vol. 405, no. 2, pp. 442 – 451, 1975.

[28] H. O. Lancaster and E. Seneta, "Chi-square distribution," *Encyclopedia of biostatistics*, vol. 2, 2005.

[29] D. Gerhard, *Audio Signal Classification: History and Current Techniques*. Citeseer, 2003.

[30] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *IEEE international conference on acoustics, speech, and signal processing*, vol. 2. IEEE, 1997, pp. 1331–1334.

[31] W. Stallings, *Cryptography and network security: principles and practice*. Pearson Upper Saddle River, 2017.

[32] Y. Jiao, L. Ji, and X. Niu, "Robust speech hashing for content authentication," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 818–821, 2009.

[33] J. P. Vilela, M. Bloch, J. Barros, and S. W. McLaughlin, "Friendly Jamming for Wireless Secrecy," in *2010 IEEE International Conference on Communications*, May 2010, pp. 1–6.

[34] W. Shen, P. Ning, X. He, and H. Dai, "Ally Friendly Jamming: How to Jam Your Enemy and Maintain Your Own Wireless Connectivity at the Same Time," in *IEEE Symposium on Security and Privacy*, May 2013, pp. 174–188.

[35] S. Haykin and Z. Chen, "The cocktail party problem," *Neural computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

[36] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[37] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," *Multichannel Speech Processing Handbook*, pp. 1065–1084, 2007.

[38] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[39] B. Zhang, Q. Zhan, S. Chen, M. Li, K. Ren, C. Wang, and D. Ma, "PriWhisper: Enabling Keyless Secure Acoustic Communication for Smartphones," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 33–45, Feb 2014.

[40] J. Scott and B. Dragovic, "Audio Location: Accurate Low-Cost Location Sensing," in *Pervasive Computing*, 2005, pp. 1–18.

[41] C. Peng, G. Shen, Y. Zhang, Y. Li, and K. Tan, "BeepBeep: A High Accuracy Acoustic Ranging System Using COTS Mobile Devices," in *Proc. of Int. Conf. on Embedded Networked Sensor Systems*, ser. SenSys '07, 2007, pp. 1–14.

[42] B. Mahafza, *Radar Systems Analysis and Design Using MATLAB*. CRC Press, 2013.

[43] P. Chiariotti, M. Martarelli, and P. Castellini, "Acoustic Beamforming for Noise Source Localization–Reviews, Methodology and Applications," *Mechanical Systems and Signal Processing*, vol. 120, pp. 422–448, 2019.

[44] T. Liu, L. Wan, Z. Qin, C. Qian, and X. Zhou, "Passive Acoustic Localization Based on COTS Mobile Devices," in *IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, 2018, pp. 299–306.

[45] J. Onaolapo, E. Mariconti, and G. Stringhini, "What Happens After You Are Pwnd: Understanding the Use of Leaked Webmail Credentials in the Wild," in *Proceedings of the 2016 Internet Measurement Conference*, ser. IMC '16, 2016, pp. 65–79.

[46] S. Kiljan, K. Simoens, D. D. Cock, M. Eekelen, and H. Vranken, "A Survey of Authentication and Communications Security in Online Banking," *ACM Comput. Surv.*, vol. 49, no. 4, pp. 61:1–61:35, Dec. 2016.

[47] M. Theofanos, S. Garfinkel, and Y. Choong, "Secure and Usable Enterprise Authentication: Lessons from the Field," *IEEE Security Privacy*, vol. 14, no. 5, pp. 14–21, Sep. 2016.

[48] R. Akhtar, Y. Shengua, Z. Zhiyu, Z. A. Khan, I. Memon, S. Ur Rehman, and S. Awan, "Content distribution and protocol design issue for mobile social networks: a survey," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, p. 128, 2019.

[49] EncapSecurity., https://www.encapsecurity.com/, 2019, accessed: 2019-04-01.

[50] Duo, Secure Two-Factor Authentication Application., https://duo.com/product/multi-factor-authentication-mfa/duo-mobile-app, 2019, accessed: 2020-01-21.

[51] Google 2-Step Verification., https://www.google.com/landing/2step/, 2019, accessed: 2020-01-21.

[52] C. S. Weir, G. Douglas, M. Carruthers, and M. Jack, "User perceptions of security, convenience and usability for ebanking authentication tokens," *Computers & Security*, vol. 28, no. 1, pp. 47 – 62, 2009.

[53] N. Gunson, D. Marshall, H. Morton, and M. Jack, "User perceptions of security and usability of single-factor and two-factor authentication in automated telephone banking," *Computers & Security*, vol. 30, no. 4, pp. 208 – 220, 2011.

[54] A. Czeskis, M. Dietz, T. Kohno, D. Wallach, and D. Balfanz, "Strengthening User Authentication Through Opportunistic Cryptographic Identity Assertions," in *Proc. of ACM Conf. on Comput. and Commun. Security*, 2012, pp. 404–414.

[55] Yubikey Neo., https://www.yubico.com/product/yubikey-5-nfc, 2019, accessed: 2020-01-21.

[56] I. Velásquez, A. Caro, and A. Rodríguez, "Authentication schemes and methods: A systematic literature review," *Information and Software Technology*, vol. 94, pp. 30 – 37, 2018.

[57] N. Karapanos, C. Marforio, C. Soriente, and S. Capkun, "Sound-Proof: Usable Two-Factor Authentication Based on Ambient Sound," in *24th USENIX Security Symposium (USENIX Security 15)*, Washington, D.C., 2015, pp. 483–498.

[58] B. Shrestha, M. Shirvanian, P. Shrestha, and N. Saxena, "The Sounds of the Phones: Dangers of Zero-Effort Second Factor Login Based on Ambient Audio," in *Proc. of ACM Conf. on Comput. and Commun. Secur.*, ser. CCS '16, 2016, pp. 908–919.

[59] P. Shrestha and N. Saxena, "Listening Watch: Wearable Two-Factor Authentication Using Speech Signals Resilient to Near-Far Attacks," in *Proc. of ACM Conf. on Secur. and Priv. in Wirel. and Mob. Netw.*, ser. WiSec '18, 2018, pp. 99–110.

[60] M. Wang, W. Zhu, S. Yan, and Q. Wang, "SoundAuth: Secure Zero-Effort Two-Factor Authentication Based on Audio Signals," in *IEEE Conference on Communications and Network Security (CNS)*, May 2018, pp. 1–9.

[61] H. T. Thu Truong, J. Toivonen, T. D. Nguyen, C. Soriente, S. Tarkoma, and N. Asokan, "DoubleEcho: Mitigating Context-Manipulation Attacks in Copresence Verification," in *IEEE International Conference on Pervasive Computing and Communications (PerCom*, March 2019, pp. 1–9.

[62] M. S. Wonsuk Choi and D. H. Lee, "Sound-Proximity: 2-Factor Authentication against Relay Attack on Passive Keyless Entry and Start System," *Journal of Advanced Transportation*, pp. 1–13, 2018.

[63] S. Drimer and S. Murdoch, "Distance bounding against smartcard relay attacks," in *Proceedings of the 16th USENIX Security Symposium", USENIX*, 2007.

[64] A. Francillon, B. Danev, and S. Capkun, "Relay Attacks on Passive Keyless Entry and Start Systems in Modern Cars," in *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2011.

[65] D. Han, Y. Chen, T. Li, R. Zhang, Y. Zhang, and T. Hedgpeth, "Proximity-Proof: Secure and Usable Mobile Two-Factor Authentication," in *Proc. of Int. Conf. on Mobile Computing and Networking*, 2018, pp. 401–415.

[66] D. Chen, N. Zhang, Z. Qin, X. Mao, Z. Qin, X. Shen, and X. Li, "S2M: A Lightweight Acoustic Fingerprints-Based Wireless Device Authentication Protocol," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 88–100, Feb. 2017.

[67] H. T. T. Truong, Xiang Gao, B. Shrestha, N. Saxena, N. Asokan, and P. Nurmi, "Comparing and fusing different sensor modalities for relay attack resistance in Zero-Interaction Authentication," in *IEEE International Conference on Pervasive Computing and Communications (PerCom)*, March 2014, pp. 163–171.

[68] H. T. T. Truong, X. Gao, B. Shrestha, N. Saxena, N. Asokan, and P. Nurmi, "Using Contextual Co-presence to Strengthen Zero-Interaction Authentication," *Pervasive Mob. Comput.*, vol. 16, no. PB, pp. 187–204, Jan. 2015.

[69] M. Fomichev, M. Maass, L. Almon, A. Molina, and M. Hollick, "Perils of Zero-Interaction Security in the Internet of Things," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 1, pp. 10:1–10:38, Mar. 2019.

[70] T. Halevi and N. Saxena, "On Pairing Constrained Wireless Devices Based on Secrecy of Auxiliary Channels: The Case of Acoustic Eavesdropping," in *Proc. of ACM Conf. on Computer and Communications Security*, 2010, pp. 97–108.

[71] W. Diffie and M. Hellman, "New Directions in Cryptography," *IEEE Trans. Inf. Theor.*, vol. 22, no. 6, Sep. 2006.

[72] S. Mirzadeh, H. Cruickshank, and R. Tafazolli, "Secure Device Pairing: A Survey," *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 17–40, First 2014.

[73] C. Soriente, G. Tsudik, and E. Uzun, "HAPADEP: Human-Assisted Pure Audio Device Pairing," in *Information Security*, 2008, pp. 385–400.

[74] M. Fomichev, F. Álvarez, D. Steinmetzer, P. Gardner-Stephen, and M. Hollick, "Survey and systematization of secure device pairing," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 517–550, 2018.

[75] M. T. Goodrich, M. Sirivianos, J. Solis, G. Tsudik, and E. Uzun, "Loud and Clear: Human-Verifiable Authentication Based on Audio," in *IEEE International Conference on Distributed Computing Systems (ICDCS'06)*, July 2006, pp. 10–15.

[76] D. Halperin, T. S. Heydt-Benjamin, B. Ransford et al., "Pacemakers and Implantable Cardiac Defibrillators: Software Radio Attacks and Zero-Power Defenses," in *IEEE Symposium on Security and Privacy*, May 2008, pp. 129–142.

[77] R. Prasad and N. Saxena, "Efficient device pairing using "human-comparable" synchronized audiovisual patterns," in *International Conference on Applied Cryptography and Network Security*. Springer, 2008, pp. 328–345.

[78] C. Peng, G. Shen, Y. Zhang, and S. Lu, "Point&Connect: intention-based device pairing for mobile phone users," in *Proceedings of the 7th international conference on Mobile systems, applications, and services*. ACM, 2009, pp. 137–150.

[79] W. R. Claycomb and D. Shin, "Secure device pairing using audio," in *43rd Annual 2009 International Carnahan Conference on Security Technology*. IEEE, 2009, pp. 77–84.

[80] M. T. Goodrich, M. Sirivianos, J. Solis, C. Soriente, G. Tsudik, and E. Uzun, "Using Audio in Secure Device Pairing," *Int. J. Secur. Netw.*, vol. 4, no. 1/2, pp. 57–68, Feb. 2009.

[81] Q. Hu and G. Hancke, "Self-jamming Audio Channels: Investigating the Feasibility of Perceiving Overshadowing Attacks," in *Radio Frequency Identification and IoT Security*. Springer International Publishing, 2017, pp. 188–203.

[82] S. Sigg, Y. Ji, N. Nguyen, and A. Huynh, "AdhocPairing: Spontaneous audio based secure device pairing for Android mobile devices," in *Proc. of Int. Worksh. on Sec. and Priv. in Spontan. Interac. and Mob. Phone Use*, vol. 12, 2012.

[83] D. Schürmann and S. Sigg, "Secure communication based on ambient audio," *IEEE Transactions on mobile computing*, vol. 12, no. 2, pp. 358–370, 2013.

[84] N. Nguyen, S. Sigg, A. Huynh, and Y. Ji, "Using ambient audio in secure mobile phone communication," in *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. IEEE, 2012, pp. 431–434.

[85] N. Nguyen and S. Sigg, "Secure Context-based Pairing for Unprecedented Devices," in *IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, Mar. 2018, pp. 119–124.

[86] M. Miettinen, N. Asokan, T. D. Nguyen, A.-R. Sadeghi, and M. Sobhani, "Context-Based Zero-Interaction Pairing and Key Evolution for Advanced Personal Devices," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '14, 2014, pp. 880–891.

[87] D. Liu, J. Chen, Q. Deng, A. Konate, and Z. Tian, "Secure pairing with wearable devices by using ambient sound and light," *Wuhan University Journal of Natural Sciences*, vol. 22, no. 4, pp. 329–336, Aug 2017.

[88] J. Han, A. J. Chung, M. K. Sinha, M. Harishankar, S. Pan, H. Y. Noh, P. Zhang, and P. Tague, "Do You Feel What I Hear? Enabling Autonomous IoT Device Pairing Using Different Sensor Types," in *2018 IEEE Symposium on Security and Privacy (SP)*, May 2018, pp. 836–852.

[89] S. Sigg, "Context-based security: State of the art, open research topics and a case study," in *Proceedings of the 5th ACM International Workshop on Context-Awareness for Self-Managing Systems*. ACM, 2011, pp. 17–23.

[90] A. Beimel, "Secret-sharing schemes: a survey," in *Int. Conf. on Coding and Cryptology*. Springer, 2011, pp. 11–46.

[91] Y. Desmedt, S. Hou, and J. Quisquater, "Audio and optical cryptography," in *International Conference on the Theory and Application of Cryptology and Information Security*. Springer, 1998, pp. 392–404.

[92] M. Ehdaie, T. Eghlidos, and M. Aref, "A novel secret sharing scheme from audio perspective," in *International Symposium on Telecommunications*. IEEE, 2008, pp. 13–18.

[93] K. Yoshida and Y. Watanabe, "Security of audio secret sharing scheme encrypting audio secrets," in *2012 International Conference for Internet Technology and Secured Transactions*. IEEE, 2012, pp. 294–295.

[94] S. Washio and Y. Watanabe, "Security of audio secret sharing scheme encrypting audio secrets with bounded shares," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7396–7400.

[95] Y. Miura and Y. Watanabe, "Security of (n, n)-threshold audio secret sharing schemes encrypting audio secrets," in *IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, 2016, pp. 1–5.

[96] M. Miettinen, N. Asokan, F. Koushanfar, T. D. Nguyen, J. Rios, A.-R. Sadeghi, M. Sobhani, and S. Yellapantula, "I Know Where You Are: Proofs of Presence Resilient to Malicious Provers," in *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security*, ser. ASIA CCS '15, 2015, pp. 567–577.

[97] J. Lopez, R. Oppliger, and G. Pernul, "Authentication and authorization infrastructures (AAIs): a comparative survey," *Computers & Security*, vol. 23, no. 7, pp. 578–590, 2004.

[98] M. Conti, N. Dragoni, and V. Lesyk, "A Survey of Man In The Middle Attacks," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 2027–2051, thirdquarter 2016.

[99] V. P. Singh and P. Pal, "Survey of different types of CAPTCHA," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 2, pp. 2242–2245, 2014.

[100] S. Sano, T. Otsuka, and H. G. Okuno, "Solving Google's Continuous Audio CAPTCHA with HMM-Based Automatic Speech Recognition," in *Advances in Information and Computer Security*. Springer Berlin Heidelberg, 2013, pp. 36–52.

[101] E. Bursztein, R. Beauxis, H. Paskov, D. Perito, C. Fabry, and J. Mitchell, "The Failure of Noise-Based Non-continuous Audio Captchas," in *IEEE Symposium on Security and Privacy*, May 2011, pp. 19–31.

[102] H. Meutzner, S. Gupta, V. Nguyen, T. Holz, and D. Kolossa, "Toward Improved Audio CAPTCHAs Based on Auditory Perception and Language Understanding," *ACM Trans. Priv. Secur.*, vol. 19, no. 4, pp. 10:1–10:31, Nov. 2016.

[103] H. Feng, K. Fawaz, and K. Shin, "Continuous Authentication for Voice Assistants," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '17, 2017, pp. 343–355.

[104] L. Blue, H. Abdullah, L. Vargas, and P. Traynor, "2MA: Verifying Voice Commands via Two Microphone Authentication," in *Proc. of Asia Conf. on Computer and Communications Security*, ser. ASIACCS '18, 2018, pp. 89–100.

[105] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "DolphinAttack: Inaudible Voice Commands," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '17, 2017, pp. 103–117.

[106] N. Maynard, "Smart homes: Vendor analysis, impact assessments & strategic opportunities 2018-2023," Juniper Research, Tech. Rep., May 2018.

[107] W. Diao, X. Liu, Z. Zhou, and K. Zhang, "Your Voice Assistant is Mine: How to Abuse Speakers to Steal Information and Control Your Phone," in *Proc. of the ACM Worksh. on Security and Privacy in Smartphones & Mobile Devices*, 2014, pp. 63–74.

[108] T. Vaidya, Y. Zhang, M. Sherr, and C. Shields, "Cocaine Noodles: Exploiting the Gap between Human and Machine Speech Recognition," in *9th USENIX Workshop on Offensive Technologies (WOOT 15)*. USENIX Association, 2015.

[109] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, "Hidden Voice Commands," in *25th USENIX Security Symposium (USENIX Security 16)*, Austin, TX, 2016, pp. 513–530.

[110] Z. Wu and H. Li, "Voice conversion and spoofing attack on speaker verification systems," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Oct. 2013, pp. 1–9.

[111] M. A. Al-Faruque, S. R. Chhetri, A. Canedo, and J. Wan, "Acoustic Side-Channel Attacks on Additive Manufacturing Systems," in *2016 ACM/IEEE 7th International Conference on Cyber-Physical Systems (ICCPS)*, Apr. 2016, pp. 1–10.

[112] D. Asonov and R. Agrawal, "Keyboard acoustic emanations," in *IEEE Symposium on Security and Privacy, 2004. Proceedings. 2004*. IEEE, 2004, pp. 3–11.

[113] L. Zhuang, F. Zhou, and J. D. Tygar, "Keyboard Acoustic Emanations Revisited," *ACM Trans. Inf. Syst. Secur.*, vol. 13, no. 1, pp. 3:1–3:26, Nov. 2009.

[114] Y. Berger, A. Wool, and A. Yeredor, "Dictionary attacks using keyboard acoustic emanations," in *Proc. of the 13th ACM Conference on Computer and Communications Security*, ser. CCS '06, 2006, pp. 245–254.

[115] A. Compagno, M. Conti, D. Lain, and G. Tsudik, "Don't skype & type!: Acoustic eavesdropping in voice-over-ip," in *Proc. of ACM Asia Conf. on Comput. and Commun. Security*. ACM, 2017, pp. 703–715.

[116] T. Halevi and N. Saxena, "Keyboard Acoustic Side Channel Attacks: Exploring Realistic and Security-sensitive Scenarios," *Int. J. Inf. Secur.*, vol. 14, no. 5, pp. 443–456, Oct. 2015.

[117] G. de Souza F. and H. Kim, "Differential audio analysis: a new side-channel attack on PIN pads," *International Journal of Information Security*, vol. 18, no. 1, pp. 73–84, 2019.

[118] T. Yu, H. Jin, and K. Nahrstedt, "WritingHacker: Audio Based Eavesdropping of Handwriting via Mobile Devices," in *Proc. of the 2016 ACM Int. Conf. on Pervasive and Ubiquitous Computing*, ser. UbiComp '16, 2016, pp. 463–473.

[119] S. A. Anand and N. Saxena, "Vibreaker: Securing Vibrational Pairing with Deliberate Acoustic Noise," in *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, ser. WiSec '16, 2016, pp. 103–108.

[120] S. A. Anand and N. Saxena, "Noisy Vibrational Pairing of IoT Devices," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–1, 2018.

[121] N. Saxena, M. B. Uddin, J. Voris, and N. Asokan, "Vibrate-to-unlock: Mobile phone assisted user authentication to multiple personal RFID tags," in *IEEE Int. Conf. on Pervasive Computing and Communications (PerCom)*, Mar. 2011, pp. 181–188.

[122] T. Halevi and N. Saxena, "Acoustic Eavesdropping Attacks on Constrained Wireless Device Pairing," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 3, pp. 563–577, Mar. 2013.

[123] Y. Kim, W. S. Lee, V. Raghunathan, N. K. Jha, and A. Raghunathan, "Vibration-based secure side channel for medical devices," in *2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, Jun. 2015, pp. 1–6.

[124] S. A. Anand and N. Saxena, "A Sound for a Sound: Mitigating Acoustic Side Channel Attacks on Password Keystrokes with Active Sounds," in *Financial Cryptography and Data Security*, 2017, pp. 346–364.

[125] P. Shrestha, S. A. Anand, and N. Saxena, "YELP: Masking Sound-based Opportunistic Attacks in Zero-effort Deauthentication," in *Proc. of ACM Conf. on Secur. and Priv. in Wirel. and Mob. Netw.*, ser. WiSec '17, 2017, pp. 195–206.

[126] S. Mare, A. M. Markham, C. Cornelius, R. Peterson, and D. Kotz, "ZEBRA: Zero-Effort Bilateral Recurring Authentication," in *2014 IEEE Symposium on Security and Privacy*, May 2014, pp. 705–720.

[127] O. Huhta, P. Shrestha, S. Udar, M. Juuti, N. Saxena, and N. Asokan, "Pitfalls in Designing Zero-Effort Deauthentication: Opportunistic Human Observation Attacks," 2015.

[128] X. Zhang, J. Liu, S. Chen, Y. Kong, and K. Ren, "PriWhisper+: An Enhanced Acoustic Short-range Communication System for Smartphones," *IEEE Internet of Things Journal*, pp. 1–1, 2018.

[129] R. Nandakumar, K. Chintalapudi, V. Padmanabhan, and R. Venkatesan, "Dhwani: Secure Peer-to-peer Acoustic NFC," in *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM*, ser. SIGCOMM '13, 2013, pp. 63–74.

[130] P. Shrestha, B. Shrestha, and N. Saxena, "Home Alone: The Insider Threat of Unattended Wearables and A Defense using Audio Proximity," in *2018 IEEE Conference on Communications and Network Security (CNS)*, May 2018, pp. 1–9.

[131] T. Li, Y. Chen, J. Sun, X. Jin, and Y. Zhang, "iLock: Immediate and Automatic Locking of Mobile Devices Against Data Theft," in *Proc. of ACM Conference on Computer and Communications Security*, ser. CCS '16, 2016, pp. 933–944.

[132] T. Halevi, D. Ma, N. Saxena, and T. Xiang, "Secure proximity detection for NFC devices based on ambient sensor data," in *European Symposium on Research in Computer Security*. Springer, 2012, pp. 379–396.

[133] A. Acar, H. Aksu, A. S. Uluagac, and M. Conti, "A Survey on Homomorphic Encryption Schemes: Theory and Implementation," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 79:1–79:35, Jul. 2018.

[134] M. Gomez-Barrero, J. Galbally, A. Morales, and J. Fierrez, "Privacy-Preserving Comparison of Variable-Length Data With Application to Biometric Template Protection," *IEEE Access*, vol. 5, pp. 8606–8619, 2017.

[135] Z. Erkin, T. Veugen, T. Toft, and R. L. Lagendijk, "Generating Private Recommendations Efficiently Using Homomorphic Encryption and Data Packing," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 1053–1066, June 2012.

[136] T. Veugen, F. Blom, S. J. A. de Hoogh, and Z. Erkin, "Secure Comparison Protocols in the Semi-Honest Model," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 7, pp. 1217–1228, Oct 2015.

[137] C. Gentry and D. Boneh, *A fully homomorphic encryption scheme*. Stanford University Stanford, 2009, vol. 20, no. 09.

[138] B. Shrestha, N. Saxena, H. Truong, and N. Asokan, "Sensor-Based Proximity Detection in the Face of Active Adversaries," *IEEE Transactions on Mobile Computing*, vol. 18, no. 2, pp. 444–457, 2019.

[139] V. S. Nagaraja, K. J. Rudresha, and S. L. Pinjare, "Design, Fabrication and Characterization of a Biologically Inspired MEMS Directional Microphone," in *2018 IEEE SENSORS*, Oct. 2018, pp. 1–4.

[140] O. Schwartz, A. David, O. Shahen-Tov, and S. Gannot, "Multi-microphone voice activity and single-talk detectors based on steered-response power output entropy," in *IEEE Int. Conf. on the Science of Electr. Engin. in Israel*, Dec 2018, pp. 1–4.

[141] M. Johansson, "VR For Your Ears: Dynamic 3D audio is key to the immersive experience," *IEEE Spectrum*, vol. 56, no. 2, pp. 24–29, Feb 2019.

Maurantonio Caprolu Maurantonio Caprolu is currently a PhD student in Computer Science and Engineering at Hamad Bin Khalifa University. He received both his Bachelor and his Master's Degree in Computer Science at Sapienza, University of Rome, Italy, on topics strictly related to applied Security and Privacy. His major research interests include security issues in Blockchain-based systems, Edge/Fog architecture and Software Defined Networking, Computer Forensics, Information Warfare.

Savio Sciancalepore Savio Sciancalepore is currently Post Doc at HBKU-CSE-ICT, Doha, Qatar. He obtained his Master's degree in Telecommunications Engineering in 2013 and the PhD in Electric and Information Engineering in 2017, both from the Politecnico di Bari, Italy. He received the prestigious award from the ERCIM Security, Trust, and Management (STM) Working Group for the best Ph.D. Thesis in Information and Network Security in 2018. His major research interests include security issues in Internet of Things (IoT) and Cyber-Physical Systems.

Roberto Di Pietro, ACM Distinguished Scientist, is Full Professor in Cybersecurity at HBKU-CSE. Previously, he was in the capacity of Global Head Security Research at Nokia Bell Labs, and Associate Professor (with tenure) of Computer Science at University of Padova, Italy. He has been working in the security field for 23+ years, leading both technology-oriented and research-focused teams in the private sector, government, and academia (MoD, United Nations HQ, EUROJUST, IAEA, WIPO). His main research interests include security and privacy for wired and wireless distributed systems (e.g. Blockchain technology, Cloud, IoT, OSNs), virtualization security, applied cryptography, computer forensics, and data science. Other than being involved in M&A of start-up—and having founded one (exited)—, he has been producing 220+ scientific papers and patents over the cited topics, has co-authored two books, edited one, and contributed to a few others. He is serving as an AE for ComCom, ComNet, Journal of Computer Security, and other Intl. journals. In 2011-2012 he was awarded a Chair of Excellence from University Carlos III, Madrid. Google says that he has been receiving 8300+ citations, with an h-index=44 and an i-index=120.