Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2019.DOI

Smartphones identification through the built-in microphones with Convolutional Neural Network

GIANMARCO BALDINI[1], (Senior Member, IEEE), IRENE AMERINI[2], (Member, IEEE)

[1]European Commission, Joint Research Centre, Ispra, 21027, Italy
[2]Dept. of Computer, Control, and Management Engineering A. Ruberti, Sapienza University of Rome, Rome, Italy

Corresponding author: Gianmarco Baldini (e-mail: gianmarco.baldini@ec.europa.eu).

The authors would like to acknowledge the European Commission through project SerIoT funded by the European Union H2020 Programme under Grant Agreement No. 780139.

ABSTRACT The use of mobile phones or smartphones has become so widespread that most people rely on them for many services and applications like sending e-mails, checking the bank account, accessing cloud platforms, health monitoring, buying on-line and many other applications where sharing sensitive data is required. As a consequence, security functions are important in the use of smartphones, especially because most of the applications require the identification and authentication of the device in mobility. This is usually achieved through cryptographic systems but recent research studies have also investigated alternative or complementary authentication mechanisms which can be used to strengthen cryptographic methods with multi-factor authentication. In this paper, we investigate the identification and the authentication of smartphones using the intrinsic physical properties of the mobile phones built-in microphones. The possibility to identify a microphone on the basis of features extracted from audio recordings is well known in literature but it is mostly used in forensics studies and usually relies on human voice recordings. On the contrary this paper proposes a smartphone identification and authentication approach by stimulating the built-in microphone with non-voice sounds at different frequencies. An extensive data set of 32 phones was used to evaluate experimentally the proposed approach. On the basis of the proven performance of deep learning techniques, a new Convolutional Neural Network architecture is proposed both for the identification and the authentication purposes. Its performance, in comparison to other machine learning algorithms, is demonstrated in presence of different types of noises (e.g., Gaussian White noise, Babble noise and Street noise). Satisfactory results have been obtained showing that the exploitation of a fingerprint from the microphone sensor is a good choice to assess smartphone distinctiveness.

INDEX TERMS smartphone identification, authentication, microphone, machine learning, deep learning

I. INTRODUCTION

The ability to identify smartphones or mobile phones (in the rest of the paper the two terms are used with the same meaning) through their built-in components has been demonstrated in the literature for various types of sensors including CCDs, accelerometers, magnetometers and also microphones. The unambiguous identification of a mobile phone can be used to perform multi-factor authentication where the physical identification related to a smartphone sensor is combined with the cryptographic authentication [1]. In such a way users of mobile applications can benefit from an improved authentication procedure in term of security and usability. This kind of identification could be extremely important also to guarantee continuous authentication for specific transactions or for time consuming processing. Furthermore, in forensics and security applications, identification proofs based on physical characteristics are much more difficult to be faked and reproduced since they are intrinsically related to the electronic component and to the mobile phone itself. So the main goal of this paper is to study and develop a methodology to identify mobile phones through the analysis of the signal coming from an on-board sensor like the microphone, with the aim to extract a smartphone fingerprint which is unambiguous and distinctive of one specific device. This identification is based on the assumption that the manufacturing process leaves some imperfections on the physical structure of each sensor, thus the output signal suffer from a systematic distortion, which is irrelevant for its use but can be distinguishable for the identification task. The extraction of features concerning the microphone sensor can contribute to the definition of the smartphone fingerprint as well as SPN (Sensor Pattern Noise) characterized the digital camera sensor of a mobile phone [2]. In a similar way, researchers have demonstrated the ability to identify mobile phones with different degrees of accuracy from various built-in MEMS sensor, such as accelerometers, radio frequency components, magnetometers and so on [3].

In particular, the microphone is a sensor that transduces acoustic pressure waves to an electrical signal. Basically, it is used in combination with a loudspeaker to allow users to communicate, digitalizing pressure waves produced by the userś voice in electric signal sequences. The microphone structure is composed by many modules and the defections in the area of the movable and conductive plate (membrane) may occur during the productive process impressing slight deviations from the ideal response of the microphone. Even assuming that such imperfections are not considerable during the standard usage of the sensors, these features may be inspected to determine the uniqueness of each microphone. In this paper and in literature (see section II), it is demonstrated that this kind of error is unique and systematic and can be used as fingerprint of the device.

The smartphone identification is related to the ability to distinguish among phones of the same model but different serial numbers (intra-model identification) and between phones of different models and brands (inter-model identification). The first one is usually more difficult to achieve because mobile phone manufacturers use the same materials to assemble the same model, while different models/brands are usually built using different components. For this reason for our experimental results we collect a dataset composed of responses from a relatively large set of mobile phones with different brands and models but with a significant number of phones belonging to the same model to better evaluate the intra-model identification capabilities.

In particular, we propose a Convolutional Neural Network method able to discriminate among various devices of the same or different brands and comparatively evaluated it against baselines like K-Nearest Neighbors (KNN), Support Vector Machine (SVM) classifiers and CNN in the presence or absence of noises. In particular, we evaluate three kind of noises: the Gaussian White noise, the Babble noise and the Street noise. In this paper, it will be demonstrated that the proposed CNN significantly outperforms related works showing a certain robustness to such noises.

The paper is organized as follows: Section II presents some previous works related to the smartphone and sensor fingerprinting. In Section III the application scenario will be described, while Section IV describes the proposed methodology and the materials used in the experimental phase. In Section V extended experimental results are presented and discussed to evaluate the performances of the proposed technique. Finally Section VI draws conclusions.

II. RELATED WORKS

Various techniques proposed so far have been devised to discern among different devices including digital cameras, smartphones, printers and scanners considering different sensors and properties.

The possibility to identify digital cameras exploiting Charge Coupled Device (CCD) sensor pattern noise has been demonstrated in well established works [4]–[6], [7]. Others papers dealing with the distinction among different kind of devices such as scanner, digital camera, computer generated content are proposed in [8]–[10]. A new trend in recent years for the device identification is related to investigate about the social networks provenance of digital images [11], [12].

The smartphone identification using built-in sensors like accelerometer, gyroscope, magnetometer is demonstrated recently by various works. The first analyzed sensor was the accelerometer in [13], then it was used in combination with the loudspeaker and the microphone [14]. The work in [15] takes into account the combination of two sensors as well (i.e., accelerometer and gyroscope). The paper in [16] presents a work on how to combine accelerometer, gyroscope and CCD sensor. In [17] an extension of the previous work is proposed where four built-in sensors are combined in order to build a more reliable fingerprint (accelerometer, gyroscope, magnetometer and microphone).

In [18] smartphone identification is used to contrast MEMS components counterfeiting using accelerometer and gyroscope, while in [19] only the magnetometer is considered.

In recent years, the problem of how to identify the source of an audio recording has been addressed, with a considerable attention to mobile phone as recording system. The advantage of using microphones for authentication in comparison to other components in the mobile phone like the CCD camera sensor or other kind of sensors is the possibility to control the stimulus, which is applied to the microphone from an external device. In this way it is easier to create a challenge/response space as described in the Section III. This is more complex for other components like a camera, where the recorded image can be random (based on the collected visual context) or for the radio frequency fingerprints where the wireless standards may impose specific constraints.

The authors in [20] proposed a pioneering work in microphone identification, where a set of audio steganalysis-based features to cluster both the microphone and the environment have been used. This work has been extended in [21], wherein a combination of statistical features and unweighted information fusion have been employed to improve the accuracy in the classification.

In most of the earliest works only the inter-model classification on speech audio recordings has been considered. More recently, in [22], [23] the authors addressed the intra-model classification task through a K-Nearest Neighbor (KNN) and Gaussian mixture model (GMM). A comparison of various features is provided showing that the use of Mel-Frequency Cepstrum Coefficient (MFCC) gives the best accuracy results in term of identification.

On the contrary, in [24] the Power Spectrum Density (PSD) of speech-free audio recordings is used to train a Support Vector Machine (SVM) classifier. The speech-free audio recordings are detected using Audacity software and the PSD is calculated using a periodogram. The authors in [25], employed MFCC coefficients of the non-speech segments of the voice recordings in combination with SVM and GMM to classify the microphones. The method exhibited promising results but it also showed substantial sensitivity to additive noises. A sparse representation of speech recording is used for device recognition in [26]. A recent work proposes to recognize different microphones based on the recorded speeches [27] using a kernel-based projection method; then again a SVM is used for the classification.

Alternatively, the microphones can be stimulated using non-voice recordings such as in [28] and [29]. In particular, the authors in [28] found out that the frequency response curve extracted from sample recordings can be a robust fingerprint to characterize the recording device. A SVM is proposed to perform the classification over 31 mobile phones. In [29], the authors proposed a speaker-to-microphone authentication protocol by leveraging the frequency response of a speaker and a microphone from two IoT wireless devices as the acoustic hardware fingerprint.

The application of deep learning on microphone identification it is a quite new task and it is inspired by the superior performance, respect to conventional machine learning methods (especially when combined with frequency representations), demonstrated for example in the radio frequency device identification (see [30] and [31]).

In [32] and [33] the authors proposed two deep learning methods to solve the microphone identification task using Convolutional Neural Networks. In particular in [32] a set of 9 devices stimulated with speech sound is employed. In [33] the proposed CNN is compared with other machine learning classifiers including SVM, Recurrent Neural Network and Random Forest. The number of microphones used in this case is 24 and a spectral representation Constant-Q Transform (CQT) is used to perform the classification. The type of sound stimuli used to generate the recordings is again related to the speech (TMIT database). The use of non-speech sounds in combination with CNN, as it is in this paper, is rather novel and it is more suitable for the identification and authentication functions that require specific sound stimuli rather than the use of speech recordings which is more appropriate for forensics analysis. To summarize, the idea introduced in this paper evaluates the use of CNN, as well as [32], [33], but a different structure of the net is proposed, non-speech audio recordings are given as input to the net and the results are tested on a superior number of phones especially to test the intra-model classification. This paper confirms the promising results shown in [32] and [33], which has proven the superior performance of CNN in comparison to conventional machine learning algorithms. As shown in Section V our experiments demonstrate the optimal behavior of the proposed CNN especially in presence of noises.

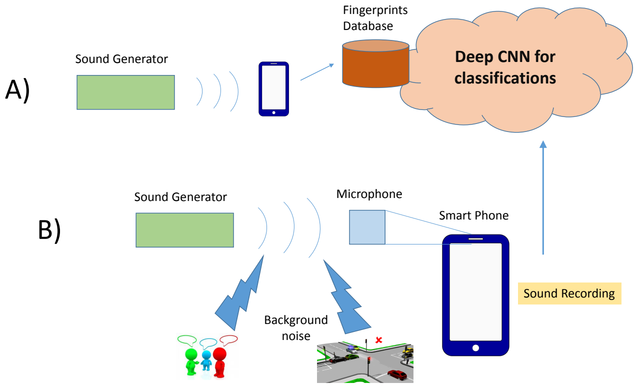


FIGURE 1: The application scenario for smartphone identification using the built-in microphone sensor.

III. APPLICATION SCENARIO

The potential application scenario for microphones identification or authentication (and consequently of the smartphone) is shown in Figure 1.

In an initial first phase (identified with A in Figure 1) the microphone of a smartphone is stimulated by a sound generator. As will be described in the Section IV the sound stimulus is composed by a repetition of audio tones. The sound recording is collected and stored in a database of fingerprints, accessible by a cloud application which is used in the in-field identification and also in the authentication phase (identified with B in Figure 1). In the second phase B, the microphone is stimulated by the same sound used in the first phase A. After that, the audio recording is sent to a deep learning cloud service able to perform the classification. In the case of the identification task the application identifies the source microphone in the fingerprints database, while in the case of the authentication task a phone with a claimed identity P_i will be compared with the fingerprint associated to P_i in order to be assigned to that phone.
In the first phase, the fingerprints are collected in a controlled sound environment avoiding the risk that the bias introduced in the recording phase become part of the fingerprint. In a real scenario the second phase is rarely ideal and background noise is certainly present. For this reason, in order to simulate this behavior, different types of noises are evaluated in the Section V-B3. In particular, three noises have been taken into account including: the Additive white Gaussian noise (AWGN), in order to consider different distances between the amplifier/loudspeaker system and the microphone, and the Babble noise and the Street noise to simulate the presence of specific background noise. Finally, the bias introduced by the loudspeaker is considered as well. We noticed that a common bias introduced by the audio amplifier (where spurious replicas at higher frequencies are generated) is mitigated in our methodology because only a segment of the frequency response is selected thus cutting those spurious replicas.

Furthermore in this application scenario, it is preferable to obtain a good identification and authentication accuracy with a limited audio recording lenght to limit the overall time for identification and the authentication. This duration time is composed by the recording of the audio stimulus in the smartphone, the transmission of the recording to the cloud application and the time for the classification itself. So the shorter is the time of the audio recording, the smaller is the size which decreases its transmission and classification times. For this reason in the Section V we perform an optimization of the size of the segment used as input to the classifier in order to reduce its length.

IV. PROPOSED METHOD AND MATERIALS

A. OVERALL WORKFLOW

The overall workflow used to generate and collect the audio recordings, process them and then submit as input to the classification procedure is presented in Figure 2.

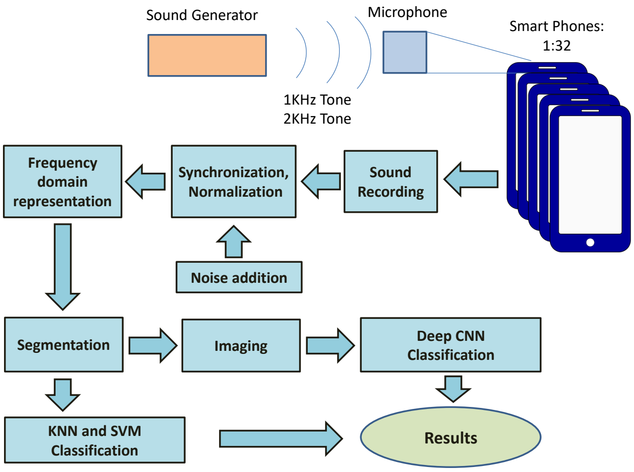


FIGURE 2: The overall methodology.

In the initial recording collection phase, each smartphone was stimulated with two separate sounds pulses at 1 KHz and 2 KHz which are repeated for 800 times (see the section below IV-C for a description on how the sounds are generated). Then, the audio recordings for each smartphone are stored in in Pulse Code Modulation (PCM) format at 44100 Hz. The audio records are then power normalized and synchronized to avoid the presence of bias related to test bed configuration (e.g., the distance from the loudspeaker or the time shift among sound recordings). Various types of noises are subsequently artificially added to the sound recordings to simulate the presence of background noise or attenuation in practical environments (see Sections V-B1 and V-B3 for details on the noise generation and related classification results). Then, the Fast Fourier Transform (FFT) is applied to the digital representation of the sound recordings to obtain a frequency representation (the Frequency Domain Representation block in Figure 2). Note that a complex time series is derived from the application of the FFT to the original sound recordings, which is expressed in real values. Even in the frequency domain, the size of the data to be classified is quite large (i.e., complex values in a frequency range from 0 to 44100 Hz) and it is necessary to perform a dimensionality reduction. After that a segmentation is performed. It was empirically found that not all the segments contribute in an equal measure to the classification. In fact, it was noticed that the best classification results was obtained by using the magnitude components of the frequency and only a specific segment was mainly responsible for the classification. This is due to the fact that most of the fingerprints are located in a frequency band between DC (direct current) and the stimulus frequency (1 KHz or 2 KHz). The empirical demonstration of this statement will be given in the Section V. This can also be seen from the amplitude of the frequency response to the stimulus at 1KHz and 2KHz, in Figures 3a) and 3b) respectively. A detail on the frequency band between 0 and 600 Hz is also shown.

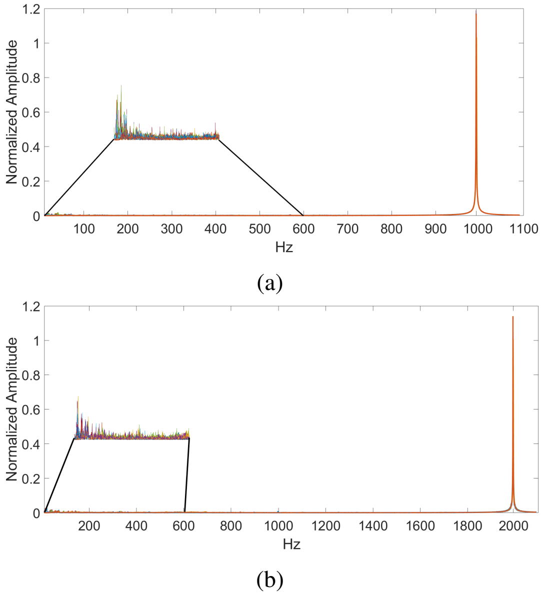


FIGURE 3: Frequency responses on the 32 mobile phones for the 1 KHz (a) and 2 KHz (b) stimulus with details on the band 0-600Hz.

Finally, the classification is performed using a Convolutional Neural Network. Two baselines classifiers (SVM and KNN) are also evaluated. A detailed description of the classification phase is provided in Section IV-B below. It is interesting to evaluate the performance of the classifier both for the inter-model identification (i.e., phones of different models and brands) and intra-model identification (i.e., phones of the same model and brand). For this reason we provide in the results Section V an analysis on the entire set of available phones to evaluate the inter-model classification performance and on a smaller set of phones of the same model to evaluate the intra-model classification.

B. THE CLASSIFICATION PHASE

The classification phase is constituted by the introduction of a Convolutional Neural Network. In the following the details related to the proposed CNN are given. In particular the scheme of the proposed architecture is shown in Figure 4 with the related optimized values. The frequency vector of the digitized microphone recording is reshaped to a matrix with different sizes according to the stimuli at 1 KHz and 2 KHz (see input layer in Figure 4). The network is then composed by three convolutional layers, the first two followed by max pooling to reduce the size. The classification performance with the use of an increasing number of convolutional layers have been evaluated too but no significant gain in classification accuracy was obtained. All the convolutional layers use the rectified linear unit (ReLU) as activation function. After that a softmax layer with as many units as the number of microphones to be identified (32 in our experiments) is attached. The softmax layer is aimed at producing the probability of each sample being classified into each class. The training phase is stopped when the loss function on the validation set reaches its minimum, at which point the model associated with a certain epoch is selected. The number of epochs for the learning rate drops is set to 10. The L2 regularization factor is set to 0.0001. To mitigate overfitting, a 4-fold approach was used for classification, where 25% of the dataset was used for test, and 75% was used for training and validation (9/10 of which used for training and 1/10 for validation, so the validation set is 7.5% of the entire dataset). The overall classification process was then repeated 20 times, each time with different training and test sets and the final results are averaged.

As introduced before, a SVM and a KNN classifiers are used in the experiments for comparison so some details regarding those baselines are provided in the following. As for CNN, to mitigate the problem of overfitting, a 4-fold approach for classification (which was repeated 20 times) was adopted.

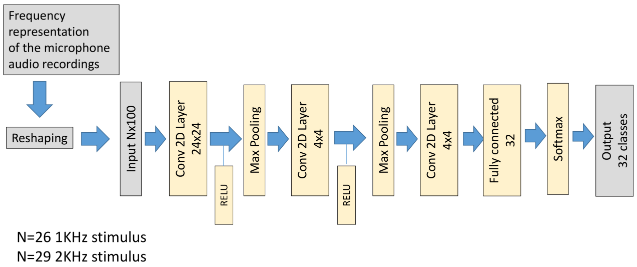


FIGURE 4: The proposed CNN architecture.

TABLE 1: List of the 32 mobile phones used in the experiments with relative IDs.

Mobile phones	ID	Quantity
Samsung ACE	1-23	23
HTC One X	24-26	3
Samsung Galaxy S5	27-29	3
Sony Experia	30-32	3
Total		**32**

In particular, the SVM was based on a Radial Basis Function (RBF) kernel and it was optimized for the values of the scaling γ and penalty C factors using a grid search approach, while the KNN uses the euclidean distance and it is optimized on the basis of the K parameter.

C. MATERIALS

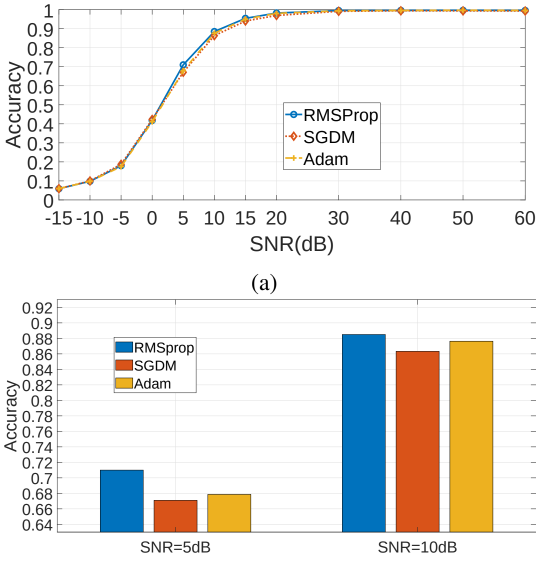
A set of 32 phones have been used to collect the audio recordings. This dataset is larger than other datasets used in the literature, and is comparable in size to the dataset recently used in [28]. The collection of smartphones used in this paper includes a larger number of phones of the same model respect to [28] to properly address the problem of the intra-model classification.

The audio signals are generated by a dedicated computer, amplified by a high quality amplifier and transmitted in the air medium with a high quality loudspeaker (to reduce the bias potentially introduced in the sound generation phase). We note that the potential impact of a lower quality of the amplifier is mitigated, in the adopted methodology, by using specific frequency bands for the audio stimuli (1KHz and 2KHz), by normalizing the audio recording and by introducing different type of noises. The microphone sensitivity and the level of the amplifier was adjusted to avoid the saturation phenomenon in the audio recordings.

The position of the phones relative to the loudspeaker is always the same in order to work in a controlled environment. Different distances of the microphone from the loudspeaker to replicate a real scenario are simulated by adding the AWGN as described in the Section V-B. On the contrary the different angle of the audio source is not considered since was demonstrated in [34] that this variation usually does not have impact in the classification performance. The smartphone was placed on a plastic absorber to minimize the effect of vibrations on the supporting surface. The audio signals are tones at 1000 Hz and 2000 Hz (1 KHz and 2 KHz) with a duration of 1 second. The audio recording was stored in the smartphone in PCM raw format at 44.1 KHz. The audio recordings were collected in different days in our laboratory in a timeframe of several weeks. The list of the mobile phones in the dataset is reported in Table 1. The total number is 32 smartphones with three different brands (Samsung, Sony and HTC) and two different model of the same brand (Samsung ACE and Samsung Galaxy S5). The quantity of smartphones for each model is reported in Table 1.

V. RESULTS

This section presents the experimental results: a) the optimization of the proposed CNN in subsection V-A and b)



(a)

(b)

FIGURE 5: Choice of the best solver function in term of accuracy in classification. a) SNR from .-15 dB to 60 dB b) the detail for SNR=5 dB and 10 dB.

the performance evaluation under different scenarios and environmental conditions in V-B.

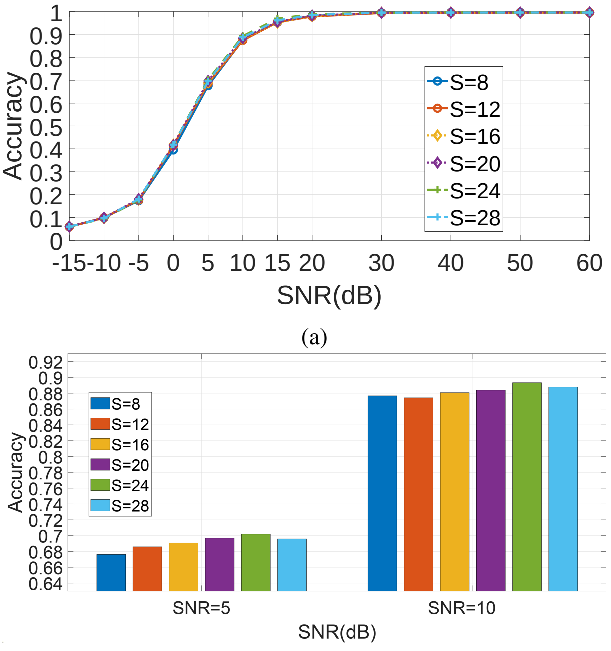
A. OPTIMIZATION

As anticipated in Section IV-B, this section will be devoted to the optimization of some of the machine learning hyper-parameters useful for the rest of the experimental analysis. In particular, the most suitable CNN optimizer and the optimal stride value related to the first convolutional layer have been considered. Different optimization algorithms have been evaluated such as the Root Mean Square Propagation (RMSProp), the Stochastic Gradient Descent with Momentum (SGDM) and Adam in term of classification accuracy varying the SNRs. In the Figure 5 a slightly improvement in the accuracy is evidenced with the RMSProp algorithm (with a decay rate of 0.999). The same behavior it is demonstrated also in Figure 5(b) where a detail for the SNR=5 and the 10 dB case is reported. Hence for this reason the RMSProp is chosen as optimization algorithm in the proposed CNN.

In Figure 6 is reported the choice of the stride values for the first convolutional layer of the CNN. Different stride values S have been evaluated $S = [8 : 28]$ with step of 4. The stride S is set to S=24 in fact from Figure in 5 (where a detail for 5 and 10 dB is reported) a small difference is appreciable. In a similar way the size of the max pooling layer was optimized to $2x2$ with a stride of 2.

For the baseline algorithm we fixed the following optimal values: the SVM grid search optimization provided values of C = 2^{12} and $\gamma = 2^6$ while the optimal value of K for the KNN was set to $K = 1$.

After the parameters of the net have been selected an analysis on the best FFT segment of the microphone recordings according to the accuracy provided in classification has been performed. In particular a two steps approach is used. In the first step, the entire frequency range of the input (1-44100 Hz) was divided in 7 segments and evaluated separately in term of accuracy. Each segment starts from the end of the previous one (e.g., 4901 Hz to 9800 Hz is the second).



(a)

(b)

FIGURE 6: Choice of the best stride value in term of accuracy in classification for the first convolutional layer. a) SNR from .-15 dB to 60 dB b) the detail for SNR=5 dB and 10 dB.

The classification result related to sound recordings at 2 KHz is shown in Figure 7. It is possible to point out that the first segment in the range of frequencies 1 - 4900 Hz provides the optimal accuracy. The segment optimization is performed at $SNR = 20$ dB.The same behavior is noticed also for other SNR values; such analysis has not been reported for the sake of conciseness.

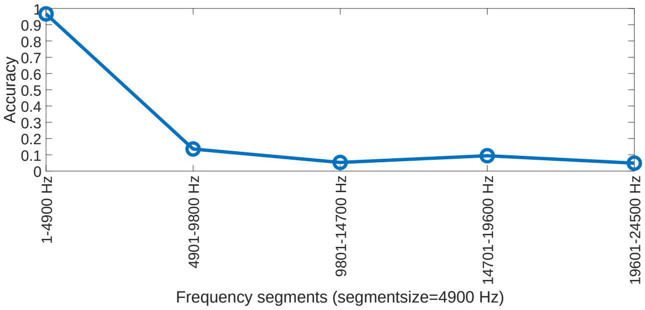


FIGURE 7: Choice of the segment in the frequency domain (2 KHz stimulus). The accuracy in classification vs the frequency segments is reported.

In the second step, the segment in the range of frequencies

1 - 4900 Hz is refined to find a smaller segment that contribute more than the others to the classification accuracy. We impose the constraint that the segment must be above 2 KHz as it is assumed that the frequency response of the stimulus at 2 KHz includes significant features of the microphone and can not be excluded. In this second step, the segments are of increasing size from a range of 1 - 2600 Hz to 1 - 5000 Hz in steps of 300 Hz. The results are reported in Figure 8, where we found out that the optimal segment is in the range 1-2900 Hz. The same procedure has also been applied to the stimuli at 1KHz, in this case the optimal frequency range is between 1-2600 Hz. So these frequency ranges have been applied in all the results provided in the following Section V-B.

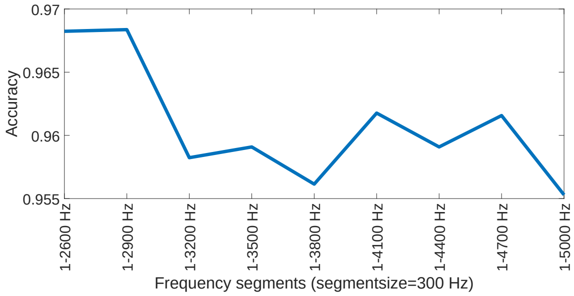


FIGURE 8: Accuracy vs frequency domain segment (2 KHz stimulus) in the range 1 - 2600 Hz to 1 - 5000 Hz.

B. PERFORMANCE EVALUATION

A comprehensive set of experiments are presented in this subsection with the aim to show the proposed method performance in relation to different operational setup, frequency stimulus and noises contamination. In particular the AWGN, the Babble and the Street noise are evaluated. The motivation behind their use is briefly given in the following: the AWGN has a constant power spectral density and it can strongly masks microphone fingerprints present in a wide range of frequencies. The Babble noise represents an unintelligible mixture of multiple speakers, which occurs frequently in our daily life and it can model a real scenario in an indoor environment. Street noise was chosen because it is a particularly noisy model and it can be used to represent an outdoor environment. The three different noises have been added digitally since is preferable that the parameters under which the data is collected is controllable to make a plausible performance assessment for an extensive range of noise magnitudes.

1) Identification in presence of AWGN

In this section, the influence of AWGN is analyzed in term of accuracy in classification. First of all we evaluate the proposed CNN method in comparison with different methods: the two baselines such as SVM and KNN and the CNN technique proposed in [33]. The optimized parameters described in section V-A are used both for the sinusoidal stimuli at 1 KHz and 2 KHz. Table 2 shows a comparison, in terms of accuracy, among the above-mentioned methods. It is evidenced that the proposed CNN method outperforms all the other methods considering both the stimulation at 1 KHz and 2 KHz also for a low SNR value (10 dB). The reported results are the median values obtained among 20 repetitions on each evaluated algorithms.

As it can be seen from Table 2, CNN performs better than other machine learning algorithms like KNN and SVM especially in presence of noise. The reason why CNN may be so effective, is because the distortions introduced by the microphone create a specific structure in the frequency domain representation of the audio signal. This structure is due to the material composition of the microphone components and to the manufacturing process, but it mostly impacts the frequency response of the microphone. This structure is not known a priori but the CNN is able extract such hidden structure and to highlight it even in presence of noise. Conventional machine learning algorithms, which operate on the basis of different principles (i.e., identification of an hyperplane as in SVM) do not produce the same classification performance since the classification model is heavily impacted by the noise.

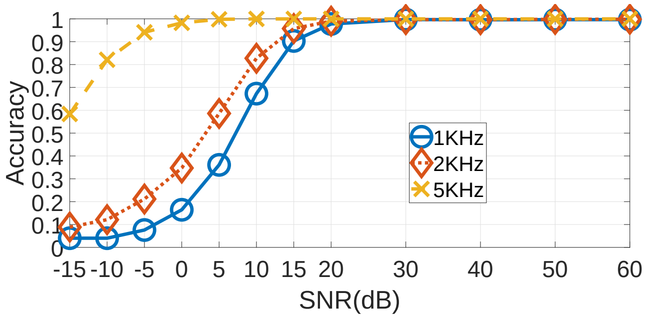


FIGURE 9: Impact of AWGN on the identification accuracy with the proposed method (the entire dataset is evaluated).

TABLE 2: Accuracy (expressed in percentage) of the proposed CNN in comparison with different methods.

Method	1 KHz	2 KHz	SNR(dB)
SVM	93.18	94.12	20
KNN	14.54	13.84	20
CNN [33]	95.01	95.30	20
CNN (proposed)	**96.00**	**96.80**	20
SVM	38.10	40.23	10
KNN	11.40	11.90	10
CNN [33]	64.75	80.90	10
CNN (proposed)	**67.27**	**82.75**	10

In the rest of the subsection the performance results of our proposed method have been analyzed considering a more extended range of Signal Noise Ratio (SNR)s from -15 to 60 dB. In particular, the classification accuracy results with the presence of AWGN for the stimuli at 1 KHz, 2 KHz and 5KHz is shown in Figure 9. It can be seen that the classification based on the stimulus at 2 KHz is more robust than the classification based on the stimulus at 1 KHz. Figure 9 also shows the performance accuracy for a stimulus at

5KHz, such behavior will be discussed later in this section. In Figure 10 the Precision and Recall metrics for the 1 Khz and 2 KHz are reported for completeness. From Figures 9 and 10 it can be also noticed that the accuracy, precision and recall degrade with lower values of SNRs, as expected, in alignment with the findings in literature [25], [28] and [29].

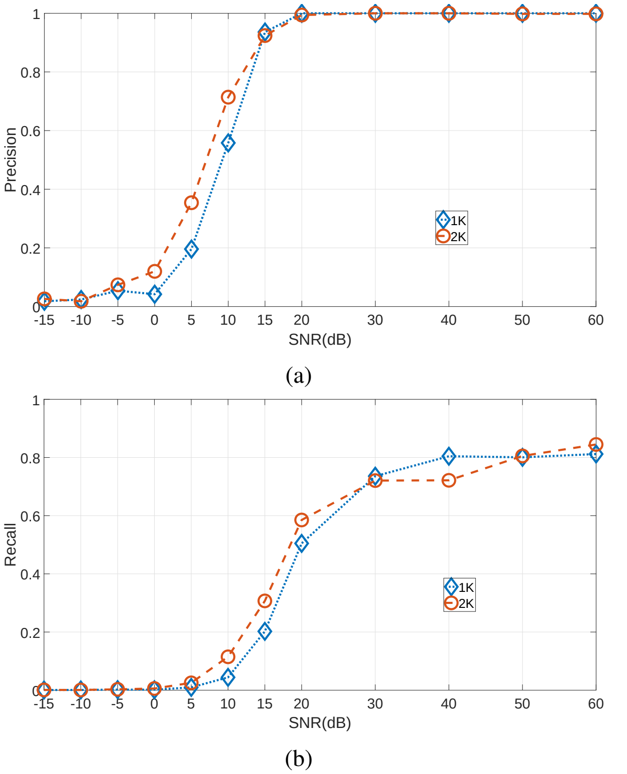
This behavior is related to the distribution of the microphone fingerprints at different frequencies and their robustness to the noise. This empirical result is useful for a practical deployment of an authentication system as it would be preferable to use stimulation at higher frequencies than at lower frequencies. On the other side, it is conceivable that the vocoding filter in the microphone introduces a fingerprint at certain frequency that is usually around 3400 Hz. Then, stimuli at higher frequencies could be more related to the presence of vocoding filter than to the fingerprint of the microphone. While the microphone fingerprint accuracy and the robustness against noise could be even higher than the stimuli at lower frequencies as evidenced in Figure 9 (5 KHz stimulus), it may not be appropriate to use them in the identification and authentication application. In fact, a change in the vocoding filter could unpair the classification process. Then, we found out that it is recommended to use stimulation at frequencies below 3400 Hz where the filter response of the phone is usually similar across mobile phones. This claim has been also demonstrated in V-A, where optimal ranges 1-2600 Hz and 1-2900 Hz are selected for the 1 KHz and the 2 KHz stimuli respectively.

Other representations of the audio recording respect to FFT has been also evaluated in the following. In fact in addition to FFT, the use of Fast Wavelet Transform (FWT) is taken into account. The results are shown in Figure 11 where the Daubechies wavelet at two different scales: 1 and 10 (DB1 and DB10 in the Figure 11) are evidenced in comparison to the FFT for the 2 KHz stimulus. It is interesting to point out that the wavelet transform is more robust than FFT for lower values of SNR (less than 5 dB) but for higher values of SNR, the identification accuracy is higher using the FFT transform. While there could be some applications where an higher accuracy at very low SNR values could be beneficial, in most practical identification and authentication applications a very high accuracy is requested. As a consequence the FFT transform is used in all the subsequent results. Similar result are obtained with the stimulus at 1 KHz and they are not presented here for brevity.

For completeness, the box plots displaying the distribution of classification accuracy are presented in Figures 12 respectively for the stimulus at 1 KHz and 2 KHz. The results are based again on a repetition of 20 times using the proposed CNN. The two diagrams show a small range of variation in the data demonstrating the robustness in classification.

2) The intra-model identification analysis

In this section the classification performance of the proposed method considering a subpart of the dataset is evaluated. In particular 23 Samsung ACE smartphones have been selected to assess the intra-model identification. The result is presented in Figure 13. It can be seen that the identification accuracy is quite high, but it is slightly lower than the intermodal identification case shown in Figure 9. Such result is expected since intra-model classification is more challenging than inter-model identification as microphones produced by the same manufacturer will share the same components like filters and amplifiers. The same behavior can be appreciated for the 2 KHz stimulus demonstrating again its superior robustness to different scenarios.



(a)

(b)

FIGURE 10: Impact of AWGN on the Precision (a) and Recall (b) using the proposed CNN.

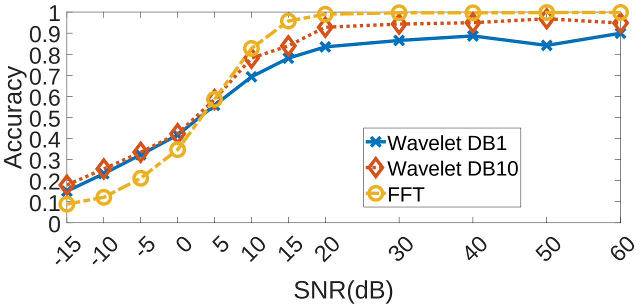
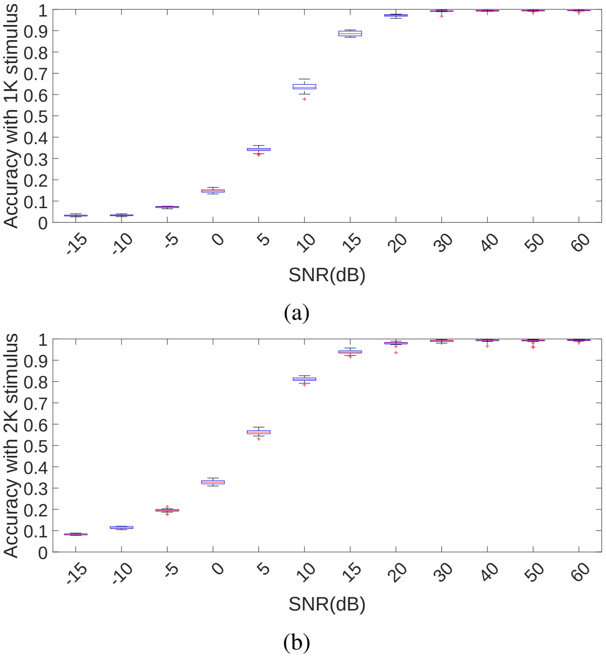


FIGURE 11: Comparison in terms of accuracy of different features (wavelets DB1, DB10 and FFT) with the proposed CNN (2 KHz stimulus).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2950859, IEEE Access



(a)

(b)

FIGURE 12: Box plot of the accuracy classification with different SNR (dB) for the 1 and 2 KHz stimulus.

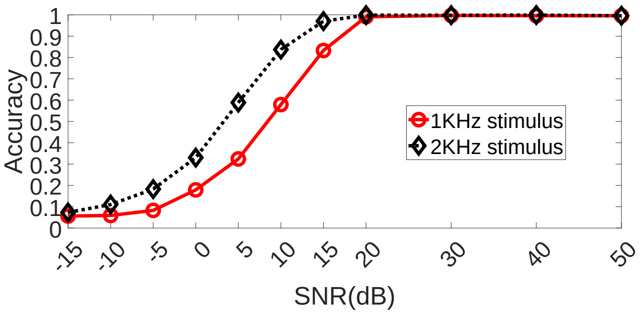
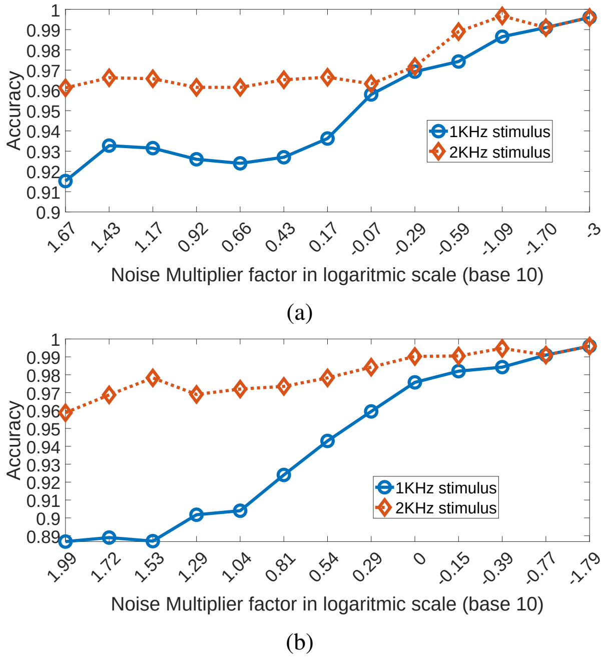


FIGURE 13: Intra-model identification versus different SNR values using the 23 *Samsung ACE* mobile phones.



(a)

(b)

FIGURE 14: Impact of Babble noise (a) and Street noise (b) on the identification accuracy with the proposed CNN on the dataset (32 phones).

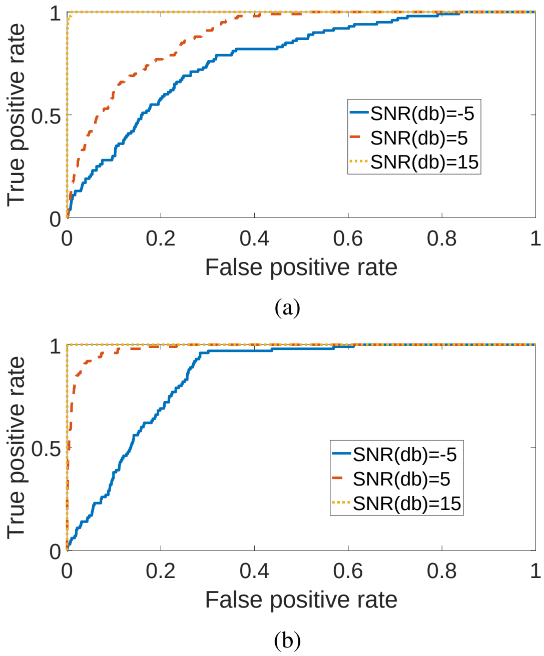
3) Identification in presence of Babble and Street noise

Different types of noise respect to AWGN have also been applied to the microphone recordings as already said in Section III. In particular the Babble noise and the Street noise, chosen from the NOIZEUS noisy speech corpus [35], are evaluated simulating the most frequent types of noises in the everyday life. In particular, the Babble noise represents an unintelligible mixture of multiple speakers and the Street noise was chosen since it is a particularly noisy model and it can be used to represent an outdoor environment. The complete dataset of 32 mobile phones are taken in account for this experiment. The results are shown in Figures 14, for increasing noise values. The x axis is in logarithmic scale (base 10) to be aligned to the white gaussian noise, which is also in logarithmic scale. As expected the presence of background noises does decrease the accuracy for both cases but without the same significant impact respect to AWGN similarly as in [33]. Again the classification performance based on the stimulus at 2 KHz is more robust than the stimulus at 1 KHz confirming the results obtained with the AWGN.

4) ROC analysis for authentication

This section will be devoted to test the authentication of a microphone. In particular, the objective is to verify that the claimed identity of a phone is confirmed or not. A potential scenario is the following: a phone B would claim the identity of a phone A. Then it is fundamental to be able to distinguish between phone A respect to any other phones which could be used to emulate it. So for the class of the phone A, a ROC (Receiver Operating Characteristic) curve is created by plotting the True Positive Ratio (TPR) against the False Positive Ratio (FPR) at various threshold settings. In particular, for the class A, the TPR is the occurrences whose the actual and the predicted values are represented by the class A, divided by the number of outputs whose predicted class is A. The FPR is the number of outputs whose actual class is not the class A, but the predicted class is the class A, divided by the number of outputs whose predicted class is not A. As in the previous experiments, the results were obtained by repeating the classification process 20 times choosing at random the training and testing sets and then averaging the results for TPR and FPR.

In particular we evaluate two cases: in the first one we chose as phone A the Samsung ACE from the intra-model dataset (with ID=3 see Table 1), in the second case, a phone from the set of Sony Experia (ID=31). In the first case (ID=3),



(a)

(b)

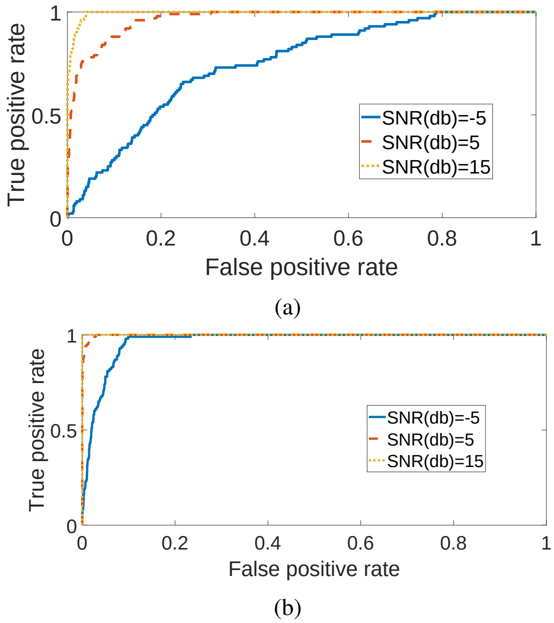
FIGURE 15: ROC curves: phone ID=3 (Samsung ACE) at different values of SNR(dB); 1 and 2 KHz stimulus.

the result is provided in Figure 15 for the stimulus at 1 and 2 KHz. In the second case (ID=31) the results are shown in Figures 16. In all cases, the results for different values of SNRs at -5, 5 and 15 dB are given.

The results obtained for the authentication confirm the outcome of the identification: the use of the stimulus at 2 KHz provides a more robust classification performance than the stimulus at 1 KHz as the ROC curves tend to be closer to the upper left corner.

For completeness, the confusion matrixes for the different stimuli and at two different values of SNR (SNR=0,10 dB) are presented hereafter. In particular, Figure 17 provides the confusion matrix at SNR=0 dB for the 1 KHz stimulus. The confusion matrix, where the number of classification errors is significant, confirms the results presented in Figure 9 where a low identification accuracy is obtained for this SNR value. Significant improvements are obtained for SNR=10 dB and 1 KHz; the related confusion matrix is shown in Figure 18. However the proposed method shows some difficulties in classifying the mobile phones with ID= 24-26 of the HTC One X and the ones from Sony Experia with ID=30-32 (i.e., in the heatmap the color of the related boxes is light blue).

Again the classification performance is higher for the 2KHz stimulus in comparison to the 1KHz stimulus. This is proven by the confusion matrixes provided in Figures 19 and Figure 20 where even for for SNR=0 dB some classes can be more easily distinguished respect to 1 KHz stimulus.



(a)

(b)

FIGURE 16: ROC curves: phone ID=31 (Sony Experia) at different values of SNR(dB); 1 and 2 KHz stimulus.

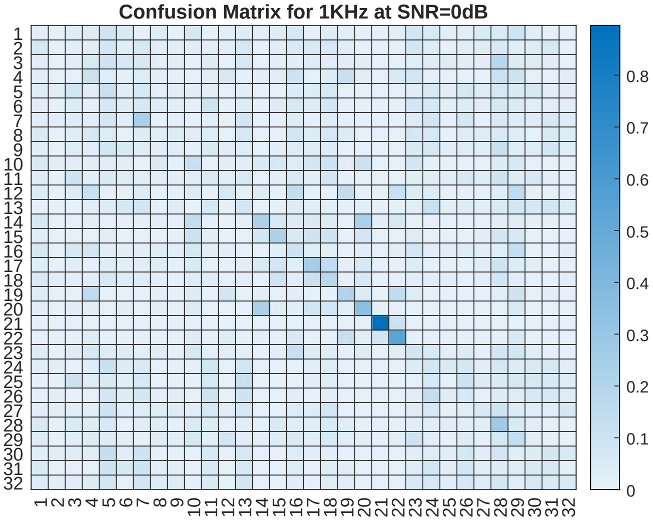


FIGURE 17: Confusion matrix on 32 smartphones; SNR=0 dB, 1KHz stimulus.

VI. CONCLUSION

In this paper, we proposed a smartphones identification and authentication method based on built-in microphones sensor for secure authentication. In our scheme, the sound registered through a microphone can be exploited in such a way that a CNN is able to learn the distinction among different smartphones. The proposed approach has demonstrated its validity using non-speech audio recordings. The experimental analysis demonstrated that the introduced CNN achieves desirable accuracy regarding both identification and authentication in various operational scenarios with different frequency stimulus and in presence of different types of noises. A comparison

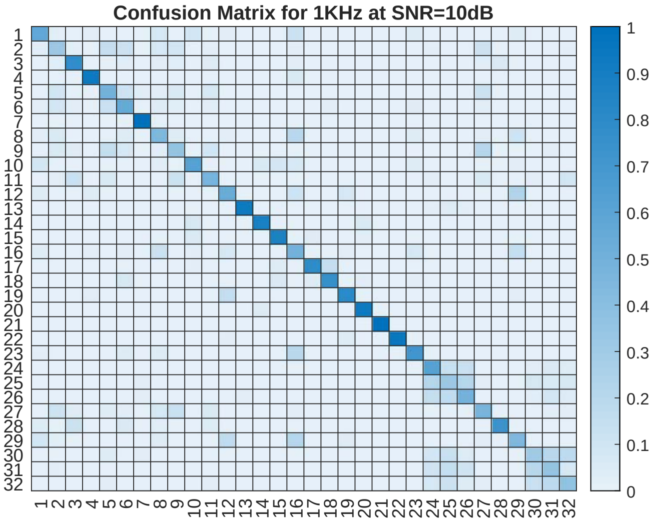


FIGURE 18: Confusion matrix on 32 smartphones; SNR=10 dB, 1KHz stimulus.

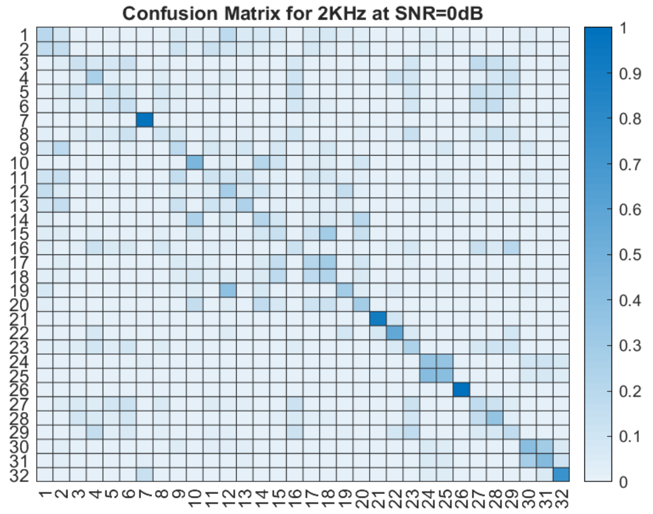


FIGURE 19: Confusion matrix on 32 smartphones; SNR=0 dB, 2KHz stimulus.

with baselines methods it is also given. Future developments will investigate the robustness of CNN against a larger set of disturbances, like reverberation effects.

REFERENCES

[1] W. C. Suski II, M. A. Temple, M. J. Mendenhall, and R. F. Mills, “Using spectral fingerprints to improve wireless network security,” in IEEE GLOBECOM 2008-2008 IEEE Global Telecommunications Conference. IEEE, 2008, pp. 1–5.

[2] J. Lukas, J. Fridrich, and M. Goljan, “Digital camera identification from sensor pattern noise,” IEEE Transactions on Information Forensics and Security, vol. 1, no. 2, pp. 205–214, June 2006.

[3] G. Baldini and G. Steri, “A survey of techniques for the identification of mobile phones using the physical fingerprints of the built-in components,” IEEE Communications Surveys & Tutorials, vol. 19, no. 3, pp. 1761–1789, 2017.

[4] M. Goljan, J. Fridrich, and T. Filler, “Managing a large database of camera fingerprints,” in SPIE Conference on Media Forensics and Security II, 2010.

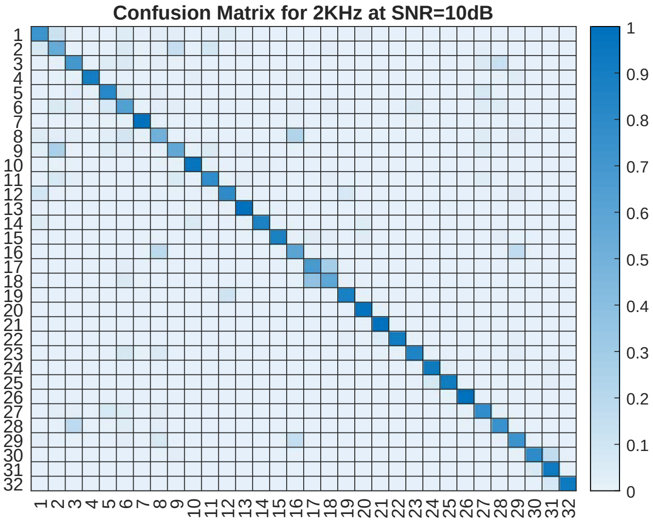


FIGURE 20: Confusion matrix on 32 smartphones; SNR=10 dB for a 2KHz stimulus.

[5] M. Goljan and J. Fridrich, “Sensor fingerprint digests for fast camera identification from geometrically distorted images,” in SPIE Conference on Media Watermarking, Security, and Forensics, 2013.

[6] I. Amerini, R. Caldelli, P. Crescenzi, A. D. Mastio, and A. Marino, “Blind image clustering based on the normalized cuts criterion for camera identification,” Signal Processing: Image Communication, vol. 29, no. 8, pp. 831 – 843, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S092359651400109X

[7] M. Chen, J. Fridrich, M. Goljan, and J. Lukas, “Source digital camcorder identification using sensor photo response non-uniformity,” in SPIE Conference on Security, Steganography, and Watermarking of Multimedia Contents, vol. 6505, 2007, pp. 65 051G–65 051G–12. [Online]. Available: http://dx.doi.org/10.1117/12.696519

[8] S. Lyu and H. Farid, “How realistic is photorealistic?” IEEE Transactions on Signal Processing, vol. 53, no. 2, pp. 845–850, 2005.

[9] N. Khanna, G. T.-C. Chiu, J. P. Allebach, and E. J. Delp, “Forensic techniques for classifying scanner, computer generated and digital camera images,” in Proc. of IEEE ICASSP, Las Vegas, USA, 2008.

[10] R. Caldelli, I. Amerini, and F. Picchioni, “A DFT-based analysis to discern between camera and scanned images,” International Journal of Digital Crime and Forensics, vol. 2, no. 1, pp. 21–29, 2010.

[11] R. Caldelli, R. Becarelli, and I. Amerini, “Image origin classification based on social network provenance,” IEEE Transactions on Information Forensics and Security, vol. 12, no. 6, pp. 1299–1308, June 2017.

[12] I. Amerini, C. Li, and R. Caldelli, “Social network identification through image classification with CNN,” IEEE Access, vol. 7, pp. 35 264–35 273, 2019.

[13] S. Dey, N. Roy, W. Xu, R. Choudhury, and S. Nelakuditi, “Accelprint: Imperfections of accelerometer make smartphones trackable,” in NDSS Symposium, 2014.

[14] H. Bojinov, Y. Michalevsky, G. Nakibly, and D. Boneh, “Mobile device identification via sensor fingerprinting,” CoRR, vol. abs/1408.1416, 2014. [Online]. Available: http://arxiv.org/abs/1408.1416

[15] A. Das, N. Borisov, and M. Caesar, “Tracking mobile web users through motion sensors: Attacks and defenses,” in 23rd Annual Network and Distributed System Security Symposium, NDSS’16, 2016.

[16] I. Amerini, P. Bestagini, L. Bondi, R. Caldelli, M. Casini, and S. Tubaro, “Robust smartphone fingerprint by mixing device sensors features for mobile strong authentication,” in Media Watermarking, Security, and Forensics. Ingenta, 2016, pp. 1–8.

[17] I. Amerini, R. Becarelli, R. Caldelli, A. Melani, and M. Niccolai, “Smartphone fingerprinting combining features of on-board sensors,” IEEE Transactions on Information Forensics and Security, vol. 12, no. 10, pp. 2457–2466, Oct 2017.

[18] G. Baldini, G. Steri, F. Dimc, R. Giuliani, and R. Kamnik, “Experimental identification of smartphones using fingerprints of built-in micro-electro

mechanical systems (MEMS)," Sensors, vol. 16, no. 6, p. 818, 2016. [Online]. Available: http://www.mdpi.com/1424-8220/16/6/818

[19] G. Baldini, G. Steri, I. Amerini, and R. Caldelli, "The identification of mobile phones through the fingerprints of their built-in magnetometer: An analysis of the portability of the fingerprints," in 2017 International Carnahan Conference on Security Technology (ICCST), Oct 2017, pp. 1–6.

[20] C. Kraetzer, A. Oermann, J. Dittmann, and A. Lang, "Digital audio forensics: a first practical evaluation on microphone and environment classification," in Proceedings of the 9th workshop on Multimedia & security. ACM, 2007, pp. 63–74.

[21] C. Kraetzer, M. Schott, and J. Dittmann, "Unweighted fusion in microphone forensics using a decision tree and linear logistic regression models," in Proceedings of the 11th ACM workshop on Multimedia and security. ACM, 2009, pp. 49–56.

[22] A. Das, N. Borisov, and M. Caesar, "Do you hear what i hear?: Fingerprinting smart devices through embedded acoustic components," in Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2014, pp. 441–452.

[23] R. Aggarwal, S. Singh, A. K. Roul, and N. Khanna, "Cellphone identification using noise estimates from recorded audio," in 2014 International Conference on Communication and Signal Processing, April 2014, pp. 1218–1222.

[24] V. Pandey, V. K. Verma, and N. Khanna, "Cell-phone identification from audio recordings using PSD of speech-free regions," in Electrical, Electronics and Computer Science (SCEECS), 2014 IEEE Students' Conference on. IEEE, 2014, pp. 1–6.

[25] C. Hanilçi and T. Kinnunen, "Source cell-phone recognition from recorded speech using non-speech segments," Digital Signal Processing, vol. 35, pp. 75–85, 2014.

[26] L. Zou, Q. He, and J. Wu, "Source cell phone verification from speech recordings using sparse representation," Digital Signal Processing, vol. 62, pp. 125 – 136, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1051200416301865

[27] Y. Jiang and F. H. F. Leung, "Source microphone recognition aided by a kernel-based projection method," IEEE Transactions on Information Forensics and Security, vol. 14, no. 11, pp. 2875–2886, Nov 2019.

[28] D. Luo, P. Korus, and J. Huang, "Band energy difference for source attribution in audio forensics," IEEE Transactions on Information Forensics and Security, vol. 13, no. 9, pp. 2179–2189, Sep. 2018.

[29] 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.

[30] K. Merchant, S. Revay, G. Stantchev, and B. Nousain, "Deep learning for RF Device Fingerprinting in Cognitive Communication Networks," IEEE Journal of Selected Topics in Signal Processing, vol. 12, no. 1, pp. 160–167, Feb 2018.

[31] G. Baldini, C. Gentile, R. Giuliani, and G. Steri, "Comparison of techniques for radiometric identification based on deep convolutional neural networks," Electronics Letters, vol. 55, no. 2, pp. 90–92, 2018.

[32] S. Qi, Z. Huang, Y. Li, and S. Shi, "Audio recording device identification based on deep learning," in 2016 IEEE International Conference on Signal and Image Processing (ICSIP). IEEE, 2016, pp. 426–431.

[33] T. Qin, R. Wang, D. Yan, and L. Lin, "Source cell-phone identification in the presence of additive noise from cqt domain," Information, vol. 9, no. 8, p. 205, 2018.

[34] C. Kraetzer, K. Qian, M. Schott, and J. Dittmann, "A context model for microphone forensics and its application in evaluations," in Media Watermarking, Security, and Forensics III, vol. 7880. International Society for Optics and Photonics, 2011, p. 78800P.

[35] Y. Hu and P. Loizou, "Noizeus: A noisy speech corpus for evaluation of speech enhancement algorithms," 2005.

GIANMARCO BALDINI Gianmarco Baldini obtained the Laurea degree in electronic engineering at the University of Rome in 1993. He has worked in the Research and Development departments of large multi-national companies in the field of wireless communications and ICT in Italy, UK, Ireland and USA before joining the Joint Research Centre (JRC) of the European Commission in 2007. In the JRC, he has been working in various areas including wireless communications, security, positioning, and machine learning and he contributed to the formulation of European policies in the areas of radio frequency spectrum, road transportation and cybersecurity. He has co-authored more than 70 peer-reviewed papers in international journals and conferences.

IRENE AMERINI (Member 17) received the Laurea degree in computer engineering and the Ph.D. degree in computer engineering, multimedia, and telecommunication from the University of Florence, Italy, in 2006 and 2010. She is currently Assistant Professor at Dept. of Computer, Control, and Management Engineering A. Ruberti, Sapienza Univeristy of Rome, Italy. She has received the Italian Habilitation for Associate Professor in telecommunications and computer science. She was a Visiting Scholar at Binghamton University, NY, USA, in 2010 and Visiting Research Fellow in 2018 at Charles Sturt University, Australia with a fellowship offered by the Australian Government Department of Education and Training, through the Endeavour Scholarship & Fellowship Program. Her main research activities include digital image processing, multimedia content security technologies, secure media, and multimedia forensics. She is a member of the IEEE Information Forensics and Security Technical Committee and EURASIP TAC Biometrics, Data Forensics, and Security. She is an Associate Editor of the IEEE ACCESS and a Guest Editor of several international journals.

• • •