CNN-based System for Speaker Independent Cell-Phone Identification from Recorded Audio

Vinay Verma and Nitin Khanna
Multimedia Analysis and Security (MANAS) Lab, Electrical Engineering,
Indian Institute of Technology Gandhinagar (IITGN), Gujarat, India,
{vinay.verma, nitin.khanna}@iitgn.ac.in

Abstract

This paper proposes a cell-phone identification system independent of speech content as well as the speaker. Audio recorded from a cell-phone contains specific signatures corresponding to that cell-phone. These unique signatures of the cell-phone implicitly captured in the recorded audio can be utilized to identify the cell-phone. These signatures of a cell-phone obtained from the recorded audio are visually more distinct in the frequency domain than in the time domain signal. Thus, by utilizing the distinctiveness of the signatures in the frequency domain and learning capability of the Convolutional Neural Network (CNN), we propose a system[1] which learns unique signatures of the cell-phones from the frequency domain representation of the audio. In particular, we have used the magnitude of the Discrete Fourier Transform (DFT) as the frequency representation of an audio signal. An extensive set of experiments performed on a large duration dataset shows that the proposed system outperforms the existing state-of-the-art systems, notably in the cases where recordings used for training and testing the systems contain mutually exclusive audio content as well as speakers.

1. Introduction

Speech is one of the most prominent ways of communication. For the last few decades, speech has been an important area of research in signal processing domain because it conveys not only the speech content itself but also various additional information, such as the language spoken, emotions, and identity of the speaker. Rapid advances and ease of availability of digital signal processing techniques in hardware as well as software packages have made digital data quite easy to acquire, process, store, and transmit. Ability to easily manipulate and tamper digital data is the flip side of the continually growing technology. Tampering digital content imposes a severe threat to the authenticity and integrity of digital data, be it digital documents, digital audio, or digital images. In the field of multimedia forensics, there are various approaches to evaluate the authenticity of digital data. Identification of brand and model of the device used to capture multimedia, referred to as source identification, is one such approach and is a crucial step during a forensic investigation. With the advancements in technology, handheld devices like cell-phones and tablets are becoming an essential part of general human life. Cell-phones are no longer used as only an instrument for making and receiving phone calls, but they serve as multipurpose hand-held devices fulfilling myriad purposes such as voice recorder and have replaced many standalone dedicated devices used for these purposes. This paper focuses on the forensics of audio recordings by identifying their originating device, cell-phone in this case. Recognizing the brand and model of a cell-phone can help in answering some of the forensic questions such as ownership verification and tampering detection in the scenarios where different parts of recorded audio are detected as originating from different cell-phones.

This paper aims to determine the exact brand and model of the cell-phone from the audio recording regardless of the speaker and speech content, in the closeset classification scenario. The proposed system for identifying the source acquisition device (cell-phone) utilizes intrinsic signatures left on the acquired audio from the acquisition device. Authors in [8], have established that the recording device leaves its own footprint on the frequency spectrum of the recorded audio and to capture these device-specific signatures from the frequency spectrum, handcrafted Mel Frequency Cepstral Coefficient (MFCC) feature vector is utilized. Authors in [18] have also observed that the frequency response

[1]Code corresponding to the proposed system is available at author's web-page.

curve of the recorded audio has the device-specific signatures and designed a feature vector named as Band Energy Difference (BED) to capture the same from the frequency response of the recorded audio signal. Our proposed system is inspired by the observations made by the previous two studies ([8] and [18]), but instead of using handcrafting on the frequency spectrum of the audio signal, we have used the magnitude of the Discrete Fourier Transform (DFT) as a simplistic frequency domain representation of the audio signal and fed to a Convolutional Neural Network (CNN) to learn the uniqueness of different cell-phones. Using this approach, the proposed system addresses one of the most common and practical scenarios of cell-phone identification where the speech content and speaker corresponding to a test recording are not present in the data available for training the system. A similar situation also arises when the test data contains multiple speakers with some of them absent in the training data, such as the recording of a telephone conversation or an interview or a meeting. This necessitates the identification system to be content as well as speaker independent. The main contributions of this paper are as follows:

1. Design of a novel CNN-based system for capturing the device signature using frequency domain representation of audio,
2. Generation of a large duration (with respect to the recording duration) dataset,
3. Better performance in speaker independent scenario, and
4. A new decision fusion approach to combine individual decisions on small audio segments to get the decisions on the given test recording.

The rest of the paper is organized as follows. Section 2 presents the relevant existing works in the literature. In Section 3, detailed description of our proposed system is given. Section 4 describes the datasets used for evaluating the performance of the proposed system. It also briefly describes the state-of-the-art methods used for comparative analysis. Experimental findings are reported in the Section 5. Finally, Section 6 concludes the paper.

2. Related Work

The earliest system for classifying microphones from the recorded audio was proposed by Kraetzer *et al.* [15]. In this paper, 7 time domain based statistical features and 56 Mel-cepstral domain features were used for classifying four microphones and maximum classification accuracies of 75.99% and 43.57% were achieved using the Bayesian classifier and K-means clustering, respectively. In [3] authors used the histogram of Fourier coefficients in the near-silence regions of the recorded audio. Maximum classification accuracy of 93.5% was achieved for classifying seven different microphones using simple logistic regression as the classifier. An alternative approach proposed in [5] used Gaussian Mixture Models (GMM) to perform close set identification of eight land-line telephones as well as eight microphones. Ikram *et al.* [10] have used polyspectral analysis for capturing the artifacts induced by the microphone. In the identification phase, distance and correlation-based similarity measures have been used. The results were evaluated on a dataset consisting of eight microphones. Hanilçi *et al.* [8] used Mel Frequency Cepstral Coefficients (MFCC) features of the recorded speech to recognize the brand and model of the cell-phone from the recorded speech. Support Vector Machine (SVM) and vector quantization have been used for classification. Their dataset consisted of 14 cell-phones of different brands and models. In the classification phase, they found that SVM does better than vector quantization and maximum closed set accuracy of 96.42% was achieved using SVM classifier. Further, Hanilçi *et al.* [6], extended their work by comparing different set of acoustic features for cell-phone classification. These feature sets were MFCC, Linear Frequency Cepstral Coefficient (LFCC), Bark Frequency Cepstral Coefficients (BFCC), and linear Prediction Cepstral Coefficients (LPCC). The study concluded that in general, the baseline MFCC does better than other features considered for comparison but with the cepstral variance normalization, LPCC performs slightly better than the MFCC. Addition of the corresponding delta features further improves the performance of the systems. In another study, Hanilçi *et al.* [7] showed that the features such as MFCC and LFCC extracted from the non-speech regions of the whole speech result in higher recognition rate for cell-phone recognition system. In another work, Pandey *et al.* [19], used power spectral density from the speech free regions of the recordings for source cell-phone classification. Authors in [1] used only the noisy part of the whole speech for MFCC feature extraction. In [14] the MFCC features were extracted from the recorded audio, and after training a GMM model, Gaussian Supervectors (GSVs) were formed using model parameters such as the mean vector and the main diagonal of the covariance matrix. Sparse representation based cell-phone verification problem has been addressed by Zou *et al.* [25]. In this work, GSVs based on MFCC features were used for building and learning different dictionaries. Deep auto-encoders were used in [17] to

extract the features and represent the intrinsic traces of a cell-phone in the speech recordings. To generate the deep representation based on the bottleneck features, a GMM model was used. Further, spectral clustering was used for associating the recordings with the cell-phone. In [23] a combination of MFCC and Inverted MFCC (IMFCC) feature vectors was utilized to emphasize the low-frequency region of the recorded audio signal along with the emphasis on the high-frequency region. The most recent system for cell-phone classification is proposed in [18], which characterizes the frequency response of a recording device using a feature descriptor named as Band Energy Difference (BED). Authors have created and tested the performance of their algorithm on a 31 cell-phone dataset in controlled conditions and a 141 cell-phone dataset in uncontrolled-conditions. Except for the recent system in [18], none of the existing systems for cell-phone classification have explicitly focused on addressing the problem in the simultaneous constraint of content as well as speaker independence. Further, the existing systems perform classification of audio segments of duration 2 seconds or larger [8,18]. The system proposed in this paper addresses these limitations of audio forensics systems by using appropriate input to the CNNs for audio source classification, which is not used by any of the existing systems.

3. Proposed System

3.1. Preprocessing

Default sampling frequencies for recording audio might differ between different cell-phones (Table 1 in Section 4) and most of the cell-phones allow users to choose from a set of audio qualities, sampling frequencies and file formats of the recorded audio. Therefore, the sampling frequency and file format of the audio recording cannot be used as a conclusive fingerprint to identify the source cell-phone from the audio recording. Thus, the proposed system is made independent of the sampling frequency and file format of the original input file by first applying a preprocessing step, where audio files are re-sampled at a fixed sampling frequency ($F_s = 8$ KHz) and saved in a lossless format (.wav). In this paper, the performance of the proposed system is evaluated on different datasets primarily containing audio recordings of human speech (Refer to Section 4 for further details of datasets). Since the telephone speech content typically ranges from 300 Hz to 3300 Hz only and general speech recognition systems use 8 KHz sampling rate for telephone speech [21]. Therefore, the pre-processing step utilized in this paper re-samples all recordings at $F_s = 8$ KHz.

3.2. Input to the Network

This paper proposes to learn the audio sensor fingerprints from the magnitude of DFT of the audio signal because: 1) sensor fingerprints utilized for forensics should be independent of speech content and learning them in frequency domain will be much easier than the time domain segmentation of spoken words, 2) the uniqueness of a device fingerprint is visually more distinctive in the frequency domain than the direct time domain representation [8], and 3) CNNs have provided high classification accuracies in many tasks related to visual discrimination of input data.

Given audio of duration T seconds (having $F_s \times T$ samples), it is divided into smaller non-overlapping segments of M_0 samples each. Each of these smaller segments corresponds to $\Delta t = M_0/F_s$ seconds. So, we have $K = \left\lfloor \frac{T}{\Delta t} \right\rfloor$ (where $\lfloor . \rfloor$ denotes the floor function) number of smaller segments of length Δt seconds. Let $\mathbf{x} \in \mathbb{R}^{M_0}$ be a smaller segment of length Δt seconds, represented in the time domain. An equivalent representation of $\mathbf{x}$ in the frequency domain can be obtained by taking M point DFT of $\mathbf{x}$. Both these representations contain the same information and are recoverable from each other as long as $M \geq M_0$. Another key property of the audio source classification system, proposed in this paper, is effectiveness on classifying small duration of audio recording and extension to forgery detection. The state-of-the-art audio source classification systems provide decisions on audio recordings of duration 2 seconds [18] or 3 seconds [8]. Thus, we have designed and tested our closed-set audio source identification system on audio recordings of duration $T = 1$ second (with $\Delta t = 0.5$ seconds). We utilize the magnitude of $M(= 8000)$ point DFT of $\mathbf{x}$. Choice of $M = 8000$ is made based on a series of initial experiments performed with different values of M. These experiments indicated that $M = 8000$ gives good trade-off between frequency resolution for audio source classification and requirement of memory and computational resources. Since the magnitude of DFT for a real signal is symmetric, we take only first 4001 $(8000/2 + 1)$ coefficients to represent $\mathbf{x}$ in the frequency domain. Let these DFT coefficients corresponding to the audio segments of Δt seconds ($\mathbf{x}$) are denoted as the vector $|\mathbf{X}|$ $\left(|\mathbf{X}| \in \mathbb{R}^{4001}\right)$. This frequency domain representation $|\mathbf{X}|$, corresponding to each of the K audio segments $\mathbf{x}$, are fed into the CNN (described later in this Section) to learn the device-specific signatures. Although some of the existing works [2, 20, 22, 24] in CNN-based systems for image forensics have utilized the histograms of the Discrete Cosine Transform (DCT) coefficients as input to the CNN, but none of the existing works in

audio forensics have utilized direct frequency domain representation as input to the CNN.

3.3. Network Architecture

Figure 1 shows the CNN architecture used in the proposed system for classifying audio segments into N classes (where N is number of cell-phones in the closed set identification problem). Input to the CNN is a 4001 dimensional DFT vector $|\mathbf{X}|$. This CNN has four 1D convolutional layers (Conv-1, Conv-2, Conv-3, and Conv-4) each having 256 filters of size 3×1 followed by max-pooling layer of size 3×1. The output of the last max-pooling layer is passed through four consecutive fully connected layers (FC-1, FC-2, FC-3, and FC-4) with 128, 128, 256 and N neurons, respectively. The ReLU activation function is used for each convolutional and first three fully connected layers. The final output is passed to a softmax function to obtain the N-class probability distribution. Batch normalization [11] is performed on the output of each convolutional layer and fully connected layers before the activation function. L2 weight regularization with a penalty parameter of 0.01 is used only with the FC-3 layer. Weights of filters in all these layers are initialized with He [9] normal initializer and bias is initialized with zero vector. Adam [13] optimizer is used with parameters β_1, β_2 and ϵ set to 0.9, 0.999, and 10^{-8}, respectively, on the batches of size 128 with categorical cross-entropy loss for optimization. CNN is trained for 30 epochs, and the best model with the lowest validation error in these thirty epochs is chosen as the final model. The learning rate is initialized with 0.001 and decayed with a factor of 10^{-1} after every ten epochs. All of our experiments are performed on an NVIDIA GeForce GTX 1080 GPU with 8 GB memory.

3.4. Decision Fusion

As we have N different cell-phones in our closed set identification problem, for each of the audio segments $\mathbf{x}$, softmax function of the CNN, gives N dimensional probability vector. Here, each of the N probability values denote the probability of that particular segment belonging to a particular class (out of N classes). We propose to use the following decision fusion scheme for predicting the audio source class of audio recordings of length T seconds. Our experiments show that this decision fusion scheme gives better performance than the simpler decision fusion schemes based on maximum probability and majority voting. For an input audio recording of duration T seconds having K consecutive audio segments of Δt seconds, we have their predicted probabilities $P_k(n)$ for each segment k ($k = 1, 2, \ldots, K$), and for each class n ($n = 1, 2, \ldots, N$). A cumulative score $\psi(n)$ for a particular class n is defined as:

$$\psi(n) = \sum_{k=1}^{K} P_k(n), \qquad \text{where } n = 1, 2, \ldots, N \tag{1}$$

Final class n^* for the audio recording of duration T seconds is obtained as:

$$n^* = \underset{n \in \{1,2,\ldots,N\}}{\arg\max}\ \psi(n) \tag{2}$$

4. Experimental Setup

MOBIPHONE dataset [14] is the only publicly accessible dataset of cell-phone recordings. This dataset consists of 21 cell-phone models of 7 different brands. Each cell-phone in this dataset has recordings of 12 male and 12 female speakers. Each speaker in the dataset has uttered 10 sentences, each sentence of approximately 3 seconds duration. In which two sentences are the same, and the other eight sentences are different for each of the speakers. Currently, for every speaker in the dataset, a single audio file of approximately $10 \times 3 = 30$ seconds is available, and each of the 10 sentences is not available separately. Thus, each cell-phone has approximately $24 \times 30 = 720$ seconds of audio recordings. While MOBIPHONE dataset originally contained 21 cell-phones, the cell-phone named as 'Samsung s5830i' (Table I in [14]) has been removed for our all experiments on MOBIPHONE dataset due to the small duration of its recordings (403 seconds).

One of the key properties of the audio source classification system proposed in this paper is robustness against the speaker and audio content changes. Therefore, for evaluating the performance of the systems addressing this problem, we will need training and testing datasets to consists of mutually exclusive audio contents and speakers. If MOBIPHONE dataset is to be used for evaluating speaker and audio content independence of systems, (after removing the two common sentences) maximum possible training size will be 9.2 minutes per class (with 23 speakers for training and remaining 1 for testing). The state-of-the-art audio source classification systems do not use CNN and have typically around 5 minutes [8], and 6 minutes [18] of training data per class. In contrast to the traditional feature engineering, CNN-based systems learn the features/signatures from the data itself and perform well when trained with a large amount of data [16]. Therefore, for training and evaluation of CNN-based system for speaker and audio content independent audio source classification, a new dataset of cell-phone recordings with larger duration for each cell-phone is required. A dataset consisting of 19 different cell-phones (Table 1)

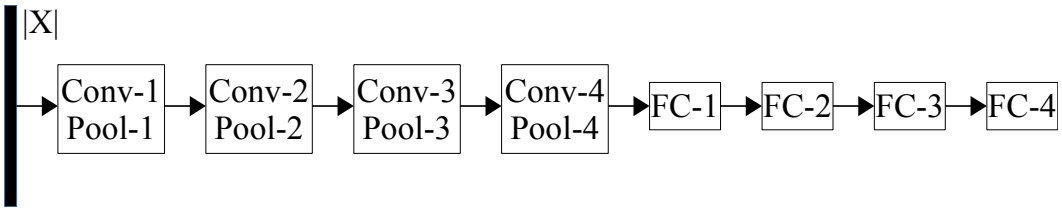


Figure 1: Network Architecture

has been prepared. This dataset is unique as compared to the publicly available MOBIPHONE dataset in terms of the duration of each recording per cell-phone. Each of the cell-phones in the database has three audio recordings S_1, S_2 and S_3, each of durations 30 minutes. The three original audio files are three webinars [2] [3] [4] with different content in English language. The three audio recordings, S_1, S_2 and S_3 belong to three different speakers, with S_1 corresponding to a female speaker while other two corresponds to two different male speakers. Cell-phones in our dataset were used as direct recording devices to record audio files being played through a loudspeaker connected to a Laptop. Audio recordings for each cell-phone were done at the same location and in a relatively quiet environment. Table 1 represents different cell-phones with their default sampling rates, and default recording formats, used in this dataset. This dataset also consists of two cell-phone (C_{18} and C_{19}) of the same brand and model. And cell-phone C_{17} is also same as C_{18}, with C_{18} having some additional features.

Based on the initial testing with different values of duration of audio recordings T and the number of non-overlapping segments K (Section 3.2), for our final system, their values are empirically chosen as $T = 1$ second and $K = 2$. Note that in all the experiments reported in this paper, the trained CNN model gives independent predictions for each of these K segments of length $\Delta t = \frac{T}{K} = 0.5$ seconds and the fusion technique described in the Section 3.4 is applied to obtain the decisions on $T = 1$ second. Hence, proposed system makes a final prediction on audio recordings of the duration $T = 1$ second, and these independent samples of $T = 1$ second are obtained by segmenting each of the three recordings into non-overlapping segments. Thus, we will get a total of $3 \times 30 \times 60 = 5400$ samples for each of the 19 classes and split them into different mutually exclusive training and testing sets for different experiments. In the rest of the paper, a sample denotes an independent audio recording of duration $T = 1$ second.

We have evaluated the performance of two state-of-the-art systems as the base-line: 1) an MFCC-based system proposed by Hanilçi *et al.* [8], and 2) a BED-based system proposed by Luo *et al.* [18]. From the implementation point of view, to calculate MFCC with the parameters described in [8], implementation available in [12] is used. Further, the Generalized Linear Discriminant Sequence (GLDS) order described in [8] is set to be 2. For the system in [18], BED feature vector are calculated by the MATLAB code provided by the authors. For the classification purpose, both systems ([8] and [18]) use LIBSVM implementation of C-SVM [4] as a classifier with a radial basis function kernel. The model parameters (C, γ) are chosen separately for each experiment using default grid search optimization available in LIBSVM on a grid of C and γ, with $C \in [-5, 15]$ and $\gamma \in [-15, 3]$. Step size used for C and γ are 3 and 2, respectively.

Table 1: Dataset Details

Sr. No.	**Class Name**	**Brand and Model**	**Default Sampling Frequency**	**Default Format**
1	C_1	HTC Desire 526G+	48.0 KHz	.3gpp
2	C_2	IPHONE 5s	44.1 KHz	.m4a
3	C_3	Lenavo PM	8.0 KHz	.amr
4	C_4	Lenovo Vibe K4 Note	48.0 KHz	.ogg
5	C_5	Moto G	16.0 KHz	.wav
6	C_6	Moto G2	16.0 KHz	.wav
7	C_7	Nokia 215	8.0 KHz	.wav
8	C_8	Nokia 311	8.0 KHz	.amr
9	C_9	Nokia Asha 201	8.0 KHz	.amr
10	C_{10}	Nokia Asha 502	8.0 KHz	.amr
11	C_{11}	Nokia 225	8.0 KHz	.wav
12	C_{12}	Samsung Galaxy Pop GT-S5570	8.0 KHz	.amr
13	C_{13}	Samsung Galaxy Grand I9082	44.1 KHz	.m4a
14	C_{14}	Samsung S5	44.1 KHz	.m4a
15	C_{15}	Micromax P680	48.0 KHz	.aac
16	C_{16}	Redmi 1s	16.0 KHz	.wav
17	C_{17}	Redmi 3s	44.1 KHz	.mp3
18	C_{18}	Redmi 3s Prime	44.1 KHz	.mp3
19	C_{19}	Redmi 3s Prime	44.1 KHz	.mp3

[2] https://in.mathworks.com/videos/managing-and-sharing-matlab-code-98671.html
[3] https://in.mathworks.com/videos/top-10-productivity-tools-in-matlab-95250.html
[4] https://in.mathworks.com/videos/spectral-analysis-with-matlab-95557.html

5. Results and Discussion

5.1. Optimal Training Duration

To find an optimal training duration, we designed an experiment where a CNN is trained with the samples obtained from first 5, 10, 15, and 20 minutes of each audio recording, S_i ($i \in \{1, 2, 3\}$). In all these four cases, testing is done on mutually exclusive samples obtained from 10 minutes duration (21 minutes to 30 minutes) of each S_i ($i \in \{1, 2, 3\}$). Note that, in the rest of the paper for the result reported on our dataset, testing with S_i ($i \in \{1, 2, 3\}$) denotes, testing the model with the $T = 1$ second samples obtained from non-overlapping segmentation of the same S_i with starting time from 21 minutes to end time of 30 minutes. For example, if all three audio recordings are used for testing, then the total duration of testing audio will be 30 minutes and corresponds to $\frac{30*60}{T} = 1800$ samples per class of duration $T = 1$ second.

Figures 2a, 2b, and 2c show the average test classification accuracies in the cases where training is done with the samples obtained from different durations of S_1, S_2, and S_3 respectively. In all three cases, when training and testing recordings belong to same speakers, accuracies are high even for 5 minutes of training data, while for other cases increasing the amount of training data leads to significant improvements in classification accuracies when training data is increased from 5 minutes to 15 minutes. Relative gain in classification accuracies is much smaller when the training data is increased from 15 minutes to 20 minutes. Thus, for the rest of the experiments presented in this paper, the total duration of training data per class is empirically fixed at 20 minutes, unless otherwise mentioned.

5.2. Effect of Audio Content and Speakers

To analyze the effect of audio content and speakers, we trained the CNN with an exhaustive combination of speakers used for training, as depicted in Table 2. As concluded from Section 5.1, total training duration is fixed to 20 minutes until specified otherwise. In Table 2, rows corresponding to S_i refer to training the CNN using first 20 minutes of that S_i. Rows corresponding to the $\{(S_i, S_j) \mid i \neq j \; ; \; i, j \in \{1, 2, 3\}\}$, refer to training with the samples obtained from the first 10 minutes of S_i, and first 10 minutes of S_j. The row corresponding to the (S_1, S_2, S_3), refers to training the CNN using first 6.7 ($\approx 20/3$) minutes from each of the three S_i. Each of the column S_i ($i \in \{1, 2, 3\}$), denotes testing with the samples obtained from 10 minutes (21 minutes to 30 minutes) duration of the S_i. Thus, in all the experiments in Table 2, testing data has different content from training data and has the same or different speaker(s) depending on the row and column number. Table 2 highlights the accuracies for experiments having the same speakers, with green color. Table 2 reveals that the average classification accuracies for intra-speaker scenarios (green color boxes) are higher than those for inter-speaker scenarios.

Table 2: Average classification accuracies (%). Rows correspond to the training audio and columns correspond to the testing audio.

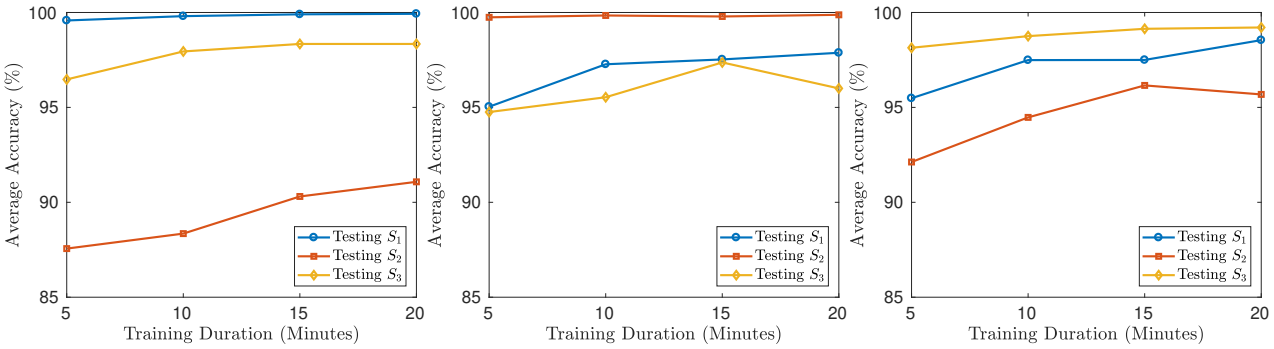
Training \ Testing	S_1	S_2	S_3
S_1	99.93	91.08	98.34
S_2	97.88	99.88	96.00
S_3	98.54	95.68	99.21
(S_1, S_2)	99.82	99.80	98.66
(S_2, S_3)	99.25	99.90	99.18
(S_3, S_1)	99.89	97.96	99.31
(S_1, S_2, S_3)	99.78	99.83	99.39

Comparison of the proposed system with systems in [8], and [18] is shown in Table 3, for the speaker dependent but content independent scenarios. This set of experiments say, E_{dep}, addresses speaker dependent scenario where training and testing data come from the same speaker with different audio content. In Table 3, each column S_i represents training with the samples from the first 20 minutes of S_i and testing with the samples from the next 10 minutes of the same S_i. Both the proposed system and the BED based system [18] perform better than MFCC based system [8], while the proposed system performs slightly better than the BED based system [18].

Table 3: Average classification accuracies (%) for experiment E_{dep}. Each column S_i represents training with the samples obtained from first 20 minutes of S_i and testing with the next 10 minutes of the same S_i.

Training	S_1	S_2	S_3
Testing	S_1	S_2	S_3
MFCC [8]	97.40	94.35	96.36
BED [18]	99.42	99.54	98.33
Proposed	99.93	99.88	99.21

Table 4 shows the comparison for the speaker as well as content independent scenarios. This set of experiments, say E_{indep}, demonstrates the speaker independent nature of the proposed system where the audio content as well as speakers, both are mutually exclusive in training and testing. In the experiments performed in E_{indep}, we train the model using samples



(a) Training with S_1 (b) Training with S_2 (c) Training with S_3

Figure 2: Average classification accuracies for different speakers and training durations.

from first 10 minutes of two audios (two out of S_1, S_2 and S_3) and test with mutually exclusive samples obtained from 10 minutes (21 minutes to 30 minutes) duration of the third audio remaining. For example, the column corresponding to (S_2, S_3) in Table 4 shows the test accuracy for classifying samples obtained from S_1 while the system is trained with samples obtained from first 10 minutes of S_2 and S_3. Table 4 clearly shows that BED based system [18] outperforms MFCC based system [8], while the proposed system significantly outperforms both of these systems. The performance gain provided by the proposed system is much higher for the speaker independent scenario (Table 4), as compared to the speaker dependent scenario (Table 3) .

Table 4: Average classification accuracies (%) in the experiment, E_{indep}. Testing with S_i denotes that the testing has been done on samples obtained from the S_i.

Training	(S_2, S_3)	(S_1, S_3)	(S_1, S_2)
Testing	S_1	S_2	S_3
MFCC [8]	65.62	64.79	79.26
BED [18]	91.87	89.61	95.24
Proposed	99.25	97.96	98.66

Detailed analysis of confusion matrices for each of these cases reveal that the proposed system not only gives higher average accuracy, but it has higher accuracy for each of the 19 classes, as compared to the BED-based system [18]. For example, confusion matrices for the proposed system and BED based system [18], in the experimental scenario E_{indep} where the systems are trained using 10 minutes audio from each of S_1 and S_2 and tested one the audio from S_3 are shown in Table 5 and Table 6, respectively. In these tables, the cells containing "−" correspond to values less than 0.005. One important aspect of this performance gain is for differentiating between the classes corresponding to cell-phones of exact same brand and model as indicated by the cluster highlighted in gray in Tables 5 and 6).

5.3. Results on Public Dataset (MOBIPHONE)

Although MOBIPHONE dataset is the only publicly available dataset of cell-phone recordings, it is not used in [8] and [18], therefore we needed to choose the recording durations and the training speakers in such a way that it closely matches with the experiments reported in [8] and [18]. We have performed two kinds of experiments, similar to E_{dep} and E_{indep} described earlier. In one experiment, similar to E_{dep}, training and testing speakers are the same, but the content is varying. Samples from the first 15 seconds from each of the 24 speakers are used for training, and remaining 10 seconds from each of the speakers are used for the testing. In the second experiment, similar to E_{indep}, content, as well as speakers both, are different. Samples from the first 25 seconds from first eight male and eight female (Table II in [14]) speakers are used for the training and initial 25 seconds from the remaining eight speakers are used for testing. Table 7, shows the average classification accuracies using the proposed system, and the systems in [8], and [18] on the MOBIPHONE dataset. The results on this publicly available dataset follow the same trend as similar experiments on our own dataset, BED based system [18] outperforms MFCC based system [8], while the proposed system significantly outperforms both of these systems. Further, the improvements provided by the proposed system are even more significant for the inter-speaker scenario.

6. Conclusion and Future Work

We have proposed a learning-based system which directly uses the frequency domain representation to extract device specific signatures from recorded au-

Table 5: Confusion matrix for the proposed system when training is done on the samples obtained from 10 minutes each of S_1 and S_2 and testing with the audio of S_3.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{17}	C_{18}	C_{19}
C_1	100.00	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C_2	-	100.00	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C_3	-	-	99.00	-	-	-	-	0.83	-	-	-	0.17	-	-	-	-	-	-	-
C_4	-	-	-	100.00	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C_5	-	-	-	-	100.00	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C_6	-	-	-	-	-	100.00	-	-	-	-	-	-	-	-	-	-	-	-	-
C_7	-	-	-	-	-	-	100.00	-	-	-	-	-	-	-	-	-	-	-	-
C_8	-	-	-	-	-	-	-	100.00	-	-	-	-	-	-	-	-	-	-	-
C_9	-	-	-	-	-	-	-	-	100.00	-	-	-	-	-	-	-	-	-	-
C_{10}	-	-	-	-	-	-	-	-	-	100.00	-	-	-	-	-	-	-	-	-
C_{11}	-	-	-	-	-	-	0.33	-	-	-	99.50	-	0.17	-	-	-	-	-	-
C_{12}	-	-	-	-	-	-	-	0.17	-	-	-	99.83	-	-	-	-	-	-	-
C_{13}	-	-	-	-	-	-	-	-	-	-	-	-	100.00	-	-	-	-	-	-
C_{14}	-	-	-	-	-	-	-	0.17	-	-	-	-	-	97.33	-	2.50	-	-	-
C_{15}	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100.00	-	-	-	-
C_{16}	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100.00	-	-	-
C_{17}	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	97.00	2.33	0.67
C_{18}	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.67	99.33	-
C_{19}	-	-	-	-	0.17	-	-	0.33	-	0.17	-	-	-	-	-	-	14.83	2.00	82.50

Table 6: Confusion matrix for the BED [18] based system when training is done on the samples obtained from 10 minutes each of S_1 and S_2 and testing with the audio of S_3.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{17}	C_{18}	C_{19}
C_1	96.83	0.50	-	0.17	0.17	-	-	1.83	-	-	-	0.17	-	-	0.33	-	-	-	-
C_2	-	98.83	-	-	-	-	-	0.67	-	-	-	0.17	-	-	0.33	-	-	-	-
C_3	-	-	97.00	-	-	-	-	2.83	-	-	-	-	-	-	-	-	-	-	0.17
C_4	0.17	0.33	0.33	96.50	0.33	0.17	-	0.17	-	-	-	-	-	-	0.50	-	-	1.50	-
C_5	-	-	-	-	98.50	1.17	-	-	-	-	-	0.17	-	-	-	-	0.17	-	-
C_6	-	-	-	-	-	99.83	-	-	-	-	-	-	-	-	-	-	-	-	0.17
C_7	-	-	-	-	-	-	99.83	-	-	-	-	-	-	-	0.17	-	-	-	-
C_8	0.17	0.83	0.17	-	0.33	0.17	-	98.17	-	-	-	-	-	-	0.17	-	-	-	-
C_9	-	-	-	-	-	-	0.17	4.67	92.83	1.67	-	-	0.50	-	0.17	-	-	-	-
C_{10}	0.17	-	-	-	-	-	-	0.17	2.17	97.50	-	-	-	-	-	-	-	-	-
C_{11}	-	-	-	-	-	-	0.67	-	-	-	97.83	-	0.50	-	-	1.00	-	-	-
C_{12}	-	0.50	0.17	-	1.17	-	-	8.00	-	-	-	90.00	-	-	0.17	-	-	-	-
C_{13}	-	-	-	-	-	-	-	1.00	3.83	0.17	-	-	94.17	-	0.33	0.50	-	-	-
C_{14}	-	-	-	-	-	-	-	1.83	0.33	-	-	-	0.50	97.00	-	0.33	-	-	-
C_{15}	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100.00	-	-	-	-
C_{16}	-	-	-	-	-	-	0.17	0.33	3.00	0.17	-	-	2.17	0.17	0.17	93.83	-	-	-
C_{17}	0.67	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	89.00	6.00	4.33
C_{18}	0.33	-	0.33	-	-	0.67	-	-	-	-	-	-	-	-	-	-	4.33	89.83	4.50
C_{19}	0.50	-	0.33	-	0.17	0.83	-	0.67	-	0.17	-	-	-	-	-	-	6.00	9.33	82.00

Table 7: Average classification accuracies (%) in the experiment E_{dep}, E_{indep} on the MOBIPHONE dataset.

	MFCC [8]	BED [18]	Proposed
E_{dep}	90.83	94.79	99.19
E_{indep}	85.50	93.70	99.30

dio. The proposed system is capable of identifying the brand and model of the cell-phones from the small duration of audio recordings. In the practical forensic scenario, where the testing audio often comes from entirely different speakers, the proposed system outperforms the current state-of-the-art methods significantly. We aim to extend the current work to 1) perform forgery detection using the proposed algorithm, that will need to have a dataset which contains both authentic and forged portion of the recordings, and 2) do automatic cell-phone identification from recordings done over the communication network. Over the network, cell-phone recordings pose additional challenge due cellular channel between speaker and recording device.

7. Acknowledgment

This material is based upon work partially supported by a grant from the Department of Science and Technology (DST), New Delhi, India, under Award Number ECR/2015/000583. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

References

[1] Aggarwal, Rachit and Singh, Shivam and Roul, Amulya Kumar and Khanna, Nitin. Cellphone Identification using Noise Estimates from Recorded Audio. In *International Conference on Communications and Signal Processing (ICCSP)*, pages 1218–1222. IEEE, 2014. 2

[2] Amerini, Irene and Uricchio, Tiberio and Ballan, Lamberto and Caldelli, Roberto. Localization of JPEG Double Compression through Multi-domain Convolutional Neural Networks. In *IEEE Conference on computer vision and pattern recognition workshops (CVPRW)*, pages 1865–1871. IEEE, 2017. 3

[3] Buchholz, Robert and Kraetzer, Christian and Dittmann, Jana. Microphone Identification using Higher-order Statistics. In *International Workshop on Information Hiding*, pages 235–246. Springer, 2009. 2

[4] Chang, Chih-Chung and Lin, Chih-Jen. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011. 5

[5] Garcia-Romero, Daniel and Espy-Wilson, Carol Y. Automatic Acquisition Device Identification from Speech Recordings. In *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 1806–1809. IEEE, 2010. 2

[6] Hanilçi, Cemal and Ertas, Figen. Optimizing Acoustic Features for Source Cell-phone Recognition using Speech Signals. In *Proceedings of the first ACM Workshop on Information Hiding and Multimedia Security*, pages 141–148. ACM, 2013. 2

[7] Hanilçi, Cemal and Kinnunen, Tomi. Source Cellphone Recognition from Recorded Speech using Non-speech Segments. *Digital Signal Processing*, 35:75–85, 2014. 2

[8] Hanilci, Cemal and Ertas, Figen and Ertas, Tuncay and Eskidere, Ömer. Recognition of Brand and Models of Cell-phones from Recorded Speech Signals. *IEEE Transactions on Information Forensics and Security*, 7(2):625–634, 2012. 1, 2, 3, 4, 5, 6, 7, 8

[9] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 4

[10] Ikram, Sohaib and Malik, Hafiz. Microphone identification using higher-order statistics. In *Audio engineering society conference: 46th international conference: audio forensics*. Audio Engineering Society, 2012. 2

[11] Ioffe, Sergey and Szegedy, Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 4

[12] Kamil Wojcicki. HTK MFCC MATLAB. https://in.mathworks.com/matlabcentral/fileexchange/32849-htk-mfcc-matlab. 5

[13] Kingma, Diederik P and Ba, Jimmy. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[14] Kotropoulos, Constantine and Samaras, Stamatios. Mobile Phone Identification using Recorded Speech Signals. In *19th International Conference on Digital Signal Processing (DSP)*, pages 586–591. IEEE, 2014. 2, 4, 7

[15] Kraetzer, Christian and Oermann, Andrea and Dittmann, Jana and Lang, Andreas. Digital Audio Forensics: A First Practical Evaluation on Microphone and Environment Classification. In *Proceedings of the 9th workshop on Multimedia & security*, pages 63–74. ACM, 2007. 2

[16] LeCun, Yann and Bengio, Yoshua and Hinton, Geoffrey. Deep learning. *nature*, 521(7553):436, 2015. 4

[17] Li, Yanxiong and Zhang, Xue and Li, Xianku and Zhang, Yuhan and Yang, Jichen and He, Qianhua. Mobile Phone Clustering From Speech Recordings Using Deep Representation and Spectral Clustering. *IEEE Transactions on Information Forensics and Security*, 13(4):965–977, 2018. 2

[18] Luo, Da and Korus, Paweł and Huang, Jiwu. Band Energy Difference for Source Attribution in Audio Forensics. *IEEE Transactions on Information Forensics and Security*, 13(9):2179–2189, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[19] Pandey, Vandana and Verma, Vicky Kumar and Khanna, Nitin. Cell-phone Identification from Audio Recordings using PSD of Speech-free Regions. In *IEEE Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pages 1–6. IEEE, 2014. 2

[20] Park, Jinseok and Cho, Donghyeon and Ahn, Wonhyuk and Lee, Heung-Kyu. Double Jpeg Detection in Mixed Jpeg Quality Factors using Deep Convolutional Neural Network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 636–652, 2018. 3

[21] Rabiner, Lawrence R and Schafer, Ronald W. *Digital Processing of Speech Signals*, volume 100. Prentice-hall Englewood Cliffs, NJ, 1978. 3

[22] Verma, Vinay and Agarwal, Nikita and Khanna, Nitin. DCT-domain Deep Convolutional Neural Networks for Multiple JPEG Compression Classification. *Signal Processing: Image Communication*, 67:22–33, 2018. 3

[23] Verma, Vinay and Khaturia, Preet and Khanna, Nitin. Cell-Phone Identification from Recompressed Audio Recordings. In *Twenty Fourth National Conference on Communications (NCC)*, pages 1–6. IEEE, 2018. 3

[24] Wang, Qing and Zhang, Rong. Double JPEG Compression Forensics based on a Convolutional Neural Network. *EURASIP Journal on Information Security*, (1):23, 2016. 3

[25] Zou, Ling and He, Qianhua and Wu, Junfeng. Source Cell Phone Verification from Speech Recordings using Sparse Representation. *Digital Signal Processing*, 62:125–136, 2017. 2