Cellphone Identification Using Noise Estimates from Recorded Audio

Rachit Aggarwal, Shivam Singh, Amulya Kumar Roul, and Nitin Khanna

***Abstract*— Rapid developments in technologies related to cell phones have resulted in their much broader usage than mere talking devices used for making and receiving phone calls. User-generated audio recordings from cell phones can be very helpful in a number of forensic applications. This paper proposes a novel system for cellphone identification from speech samples recorded using the cellphone. The proposed system uses features based on estimates of noise associated with recordings and classifies them using sequential minimal optimization (SMO) based Support vector machine (SVM). The performance of the proposed system is tested on a custom database of twenty-six cell phones of five different manufacturers. The proposed system shows promising results with average classification accuracy around 90% for classifying cell phones belonging to five different manufacturers. The average classification accuracy reduces when all the cell phones belong to the same manufacturer.**

***Index Terms*— Audio forensics, Cellphone identification, Mel-frequency cepstrum coefficient (MFCC), Multimedia source identification, Noise estimation.**

I. Introduction

THE present world is living in an era of digital domain wherein storage, modification, and replication of any textual/audio/visual information is easy and rapid. The sophistication and, ease of availability and usability of software for editing any digital media has made it very easy even for a novice to make realistic modifications to the digital media. This gave rise to a new problem because it has become hard even for a well-trained examiner to detect different varieties of digital media manipulations/frauds. The field of digital audio forensics is dedicated to detecting such threats and frauds for speech/audio signals. Present audio forensics techniques make use of digital signal processing for detecting the authenticity of recorded speech signal, identifying talkers, improving speech intelligibility, and interpreting evidences. This paper deals with rather new problem in this area, to detect the effect of mobile-phone voice recorder on processing of the signal recorded by that device. This research area is motivated by the fact that many of the crimes these days involve significant evidences generated from mobile phones that can become important evidence and can also be presented in court-of-law.

For different speech processing applications, speech signal is processed in an appropriate way to extract relevant features that are specific to the application area concerned. The mobile-phone identification system also aims to extract features from the speech signals that represent characteristics of the mobile-phone such as its transfer function and then classify the mobile phones based on the extracted features. Various methods used for feature extraction from the recorded audio signals include Mel-Frequency Cepstrum Coefficient (MFCC), Perceptual Linear Prediction (PLP), Bark-Frequency Cepstrum Coefficient (BFCC), and Linear Predictive Coding (LPC) [1]. These features are then classified using different classifiers such as support vector machine (SVM), Gaussian Mixture Model (GMM), and vector quantization [2].

This paper proposes a novel system for cellphone identification from speech samples recorded using the cellphone. The feature extraction phase of the proposed system involves two steps. First step is the estimation of noise samples corresponding to the recorded speech. The second step of feature extraction is to estimate MFCC features for these estimated noise samples. In the classification step, first the large numbers of extracted MFCC feature vectors are clustered using k-means clustering. Then the feature vectors corresponding to the centroid of these clusters are classified using SMO based SVM classifier. The rest of the paper is organized as follows. Section II consists of literature review of different methods for cellphone recognition. Section III covers the details of the proposed system followed by the results in Section IV and finally conclusions and future work are given in Section V.

II. Literature Survey

A number of practical attempts have been made for identifying recording devices or microphones. An initial attempt to determine microphone and recording environment was made in [3]. The experiment was conducted on a set of four microphones. Ten different recording environments were selected and ten audio files per environment were recorded

Rachit Aggarwal is a part time M. Tech. student in the Department of Electronics and Communication, Graphic Era University and Lecturer in the Department of Electronics and Communication Engineering, Uttaranchal University, Dehradun. Shivam Singh is a B.Tech student in the Department of Electronics and Communication, Graphic Era University (e-mail: rachit.edu@gmail.com, vishi.singh91@gmail.com).

Amulya Kumar Roul and Dr. Nitin Khanna (corresponding author) are Assistant Professor and Associate Professor, respectively in the Department of Electronics and Communication Engineering, Graphic Era University, Dehradun, Uttarakhand- 248002, India (e-mail: roul.amulya@gmail.com, dr.nitin.khanna@alumni.purdue.edu).

978-1-4799-3358-7/14/$31.00 ©2014 IEEE

from each of the four microphones at 44.1 KHz. The experiment was aimed at evaluating three hypothesis i.e. possibility of correctly classifying the microphones, possibility of correctly classifying recording environment and the effect of feature selection on accuracy improvement. In that system the features of speech were extracted based on audio steganalysis. These features are then classified using k-means and Naive Bayes classifier. The method proposed in [3] gives 75% accuracy for microphone classification and 41% accuracy for room identification. The experiments also showed that reducing the number of features lead to reduction in classification accuracy.

Motivated by the application of device identification with audio recordings, another Fourier transform based method was proposed in [4]. They extracted Fourier coefficients using 512 and 4096-point FFT on non-overlapping speech frames of near silence regions using 9 different thresholds for noise/silence frame identification. Individual feature vectors from each of these frames are summed to form the feature vectors for each of the recording and then classified using different classifiers available in Weka [5]. This method [4] gave average classification accuracy of 93% with Simple Logistic Classifier for classifying four microphones.

A three level information fusion-match, rank and decision approach for optimally combining the decisions at various levels of decision tree and linear logistic regression models was proposed in [6]. This fusion approach showed improvement in classification results for rank- and decision-level fusion with classification approaching maximum100% in certain cases as opposed to previous works in [3,4]. Further analysis of reliability of these classification methods is yet to be done.

A system based on the usage of Gaussian super vectors for frequency domain information characterization is proposed in [7]. This system uses MFCCs and Linear Frequency Cepstrum Coefficients (LFCCs) for parameterization with SVM as the classifier. The experiments reported in [7] show an average classification accuracy of around 90% for two different sets: landline and microphone recordings. These experiments accounted for variability in terms of test record duration, phone sets and parameterizations.

The cellphone identification system proposed in [8] modeled the complete transfer function corresponding to a cellphone recorder as multiplication of transfer function of phone and the vocal tract excitation function. This approximation of transfer functions reduced the problem to speaker identification problem. The identification is then based on MFCC feature extraction of speech along with their delta coefficients to identify different phones. For closed-set identification on 14 different cell-phones, this system [8] gave an average classification accuracy of around 93% and 96% using vector quantization and SVM, respectively.

A random spectral feature based approach to extract feature from each recording device is performed in [9] where the proposed method is applied on eight landline telephones of Lincoln-Labs Handset Database. For comparison linear SVM and nearest neighbor classifiers were used that outperformed MFCC on any classifier with an accuracy of 95.55%.

Another recent work presented comparison of classification accuracy using MFCC, Linear Predictive Cepstrum Coefficient (LPCC) and Perceptually-based Linear Predictive Coefficients (PLPC) as feature extraction methods with GMM used as a classifier. They also evaluated the effect of duration of test and training signals. Sixteen different microphones were tested on three databases with one containing same speaker same content, another containing different recorded data and last was a subset of TIMIT database. Classification results show that LPCC performed better than MFCC and PLPC with accuracy results approaching 100% for GMM mixture size value > 16 for test data duration of 3 sec [10].

A blind-passive approach for handset identification [11] involved extracting the sketches of spectral features by averaging its spectrogram along time axis and then mapping mean spectrogram into a low dimensional space. Thus, in presence of sufficient speech recordings, the Sketches of Spectral Features SSFs extracted from the recordings are later used to classify the test sets using sparse-representation based classifier. Experiments performed on eight telephone handsets from Lincoln-Labs Handset Database yielded an accuracy of around 95%.

III. Proposed System

Different systems proposed in existing literature varied in terms of feature extraction and different classifier schemes. However, most of them extracted features either directly from the recorded speech signals or from the trimmed version of signal that is selected based on comparing the speech signal with some pre-selected threshold. They did not concentrate on pre-processing the speech signal to either directly predict the transfer function of the cell-phone or extract features from a modified signal that more directly depicts the effect of the cell-phone's transfer function.

Fig. 1 shows an overview of the system proposed in this paper. The proposed system processes the original recorded signal to extract out the noise spectrum corresponding to the input speech signal.

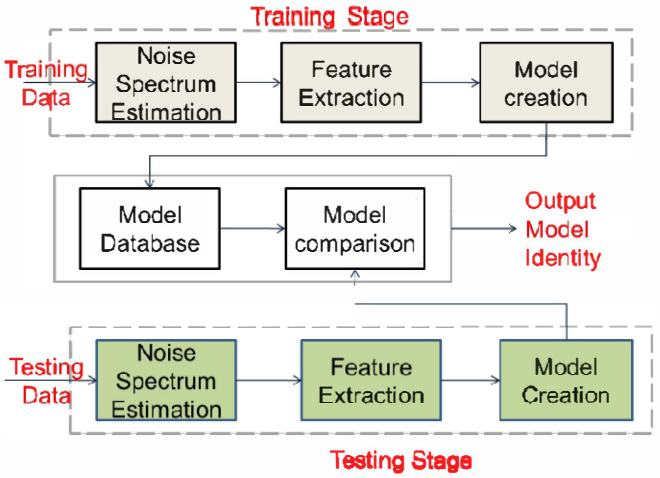

Fig. 1. Overview of the Proposed System for Speech Recorder Identification.

The purpose of extracting noise is to eliminate the contribution of speech/voice in the signal so that the mobile recorder's transfer function can be more effectively captured without the interference of voiced spectrum. The noise spectrum is extracted using a multi-band spectral subtraction method proposed earlier for enhancing speech corrupted by colored noise [12]. This method [12] of noise estimation considered noise to be additive and uncorrelated with clean signal. Let $r(n)$, $s(n)$ and $d(n)$ denote the noisy input signal recorded by the cellphone, clean speech signal and corresponding noise sequence respectively. Then Equation 1 shows the noise model used for noise estimation.

$$r(n) = s(n) + d(n) \quad (1)$$

The spectrum of corrupted signal can then be given as:

$$R_i(k) = S_i(k) + \widehat{D}_i(k) \quad (2)$$

where $R_i(k)$, $S_i(k)$and $\widehat{D}_i(k)$ represents short-time Fourier transform (STFT) of noisy input, clean and estimated noise signal respectively of the i^{th} frame. The system aims at predicting estimated colored noise $\widehat{D}_i(k)$ from the signal spectrum by dividing the spectrum into frequency bands and then estimating the noise for each band. Initially, a fixed length of initial segment of speech input is treated as noise and its mean spectrum is computed. The mean spectrum is updated for each frame where the frame classified as speech absent frame contributes more in updation of noise spectrum. The resulting noise spectrum combined with other parameters is subtracted from noisy speech spectrum to obtain clean speech spectrum [12]. The resulting noise spectrum estimate over a frequency band for each frame can be computed as:

$$N_j(k) = R_j(k) - S_j(k) \;\; b_j < k < a_j \quad (3)$$

where b_j and a_j are beginning and end frequency bins of j^{th} frequency band. The above equation subtracts clean speech spectrum by noisy input spectrum over different frequency bands to obtain the noise spectrum estimate. The resulting spectrum is colored noise spectrum estimate from input speech.

The estimated noise signal is fed to feature extraction stage. Our proposed system uses mel-frequency cepstrum coefficient (MFCC) for extracting features because of its wide use in speech processing. MFCC divides the signal into number of frames of short duration that are used as inputs to filterbank stage after taking magnitude squared Fourier transform sequence. The filterbank consists of filters or frequency bands that are equally spaced in the mel-scale of frequency (Fig. 2).

The MFCC coefficients can be obtained by [1, 8]:

$$C_m = \sum_{n=1}^{M} [\log (H(m)] \cos \left[\frac{\pi m}{M}\left(m - \frac{1}{2}\right)\right] \quad (4)$$

C_m is the m$^{\text{th}}$ MFCC coefficient computed by taking the DCT of log output of individual filter H(m) with M filters in the mel-frequency filterbank. The required ten to twelve coefficients are selected from these MFCC coefficients (C_m). The delta coefficients d_t for a particular frame are obtained using the following equation [1, 8]:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \quad (5)$$

For capturing the dynamics of MFCC coefficients, delta coefficients are estimated over a number of frames.

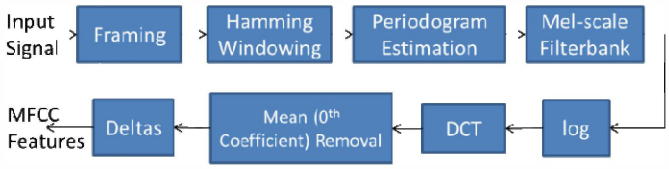

Fig. 2. Steps Involved in Extraction of MFCC Features.

Typically, N, which depicts number of frames to be used for delta computation, is varied from two to five. Block diagram implementation of this stage is shown in Fig. 2. The MFCC and their delta feature vectors extracted from individual frame of a speech signal form a very large matrix of feature vectors. Therefore, to reduce the number of feature vectors for each audio file, k-means clustering is done with k being the number of clusters to be extracted. K-means is one of the simplest unsupervised learning algorithms that aim to partition *N* feature vectors into *k* clusters such that each feature vector is identified by a cluster to which it is closest. The approximation function performs distance minimization between data-points and cluster centers. Also each cluster center should be placed so as to keep inter-cluster distance to maximum.

The proposed system shown in Fig. 1 consists of two phases. Training phase trains the system by extracting feature sequence from training signals with known source and saving the generated models. Testing phase takes test signal of unknown origin as an input and compares its feature vectors to the saved models in the database. The output of model comparison stage tells the predicted source cell-phone for a particular input test speech and can be used to form the confusion matrix.

Confusion matrix depicts actual instances along the rows and predicted class instances along the columns. The diagonal elements of such matrix correspond to percentage of correctly classified instances for a particular class. Confusion matrix provides an easy way to conclude whether (and how much) the system is confusing different classes with each other. The result in the form of confusion matrix is generated using SVM as a classifier with sequential minimal optimization (SMO) algorithm. For the experiments presented in this paper, we utilized Weka [5], open source software for data mining.

IV. Results and Discussions

A. Speech Database

A custom database of randomly selected 26 cell phones of different brands as well as different sets of same make and model was created by recording different speech content. Recording of approximately 15 min per cell phone was done in a uniform environment including welcoming note, seven days of week, names of months, numbers from zero to 20 repeated three times each, domain specific information comprising of 5 min speech and an introductory speech of 1 to 2 min. The numbers of cellphones selected were to perform experiment on a larger set compared to previous experiments. Since our aim is to evaluate systems for cellphone

identification and not for speaker or speech content identification. Therefore, the primary goal of the gathered database is to capture variations due to cellphones, keeping other factors fixed. All the files are recorded by the same speaker, except the last file of introduction that is recorded by different owners of the phones. The in-built voice recorder of these cellphones records speech in AMR format which is then converted into WAV format with the same specifications such as same sampling rate, using FFMPEG library [13]. Table 1 shows list of cell-phones used for experiments. For the experimental results presented in this paper, recordings of names of days and months are used. For each of these two categories, three audio files per category are available. From the extracted feature vectors, training and test sets are configured and performance results are obtained for different train and test setups.

Initially, speech was hamming windowed into 25 ms frames with 15 ms overlap and STFT of frame was taken to predict the noise spectrum that is later used to extract 12 MFCC coefficients from each frame along with their delta coefficients resulting in 24 coefficients per frame. K-means clustering with different codebook sizes k= 8, 16, 32, 64 and 128 is performed on MFCC feature vectors of each audio file to obtain a uniform and compact feature matrix. The dataset is randomly split into two non-overlapping subsets to be used for training and testing of the proposed system.

Performance of the proposed system based on noise spectrum estimation (referred as system based on estimated noise features, ENF) is compared with the system that extracts features directly from the recorded speech without doing any noise estimation (referred as system based on speech features, OSF). We will refer these two methods as noise-estimation based features and direct-speech features.

B. Classification on Fixed Speech Content

The first experiment is aimed to evaluate the performance of the proposed system when training and testing is performed on the speech with similar content. Leave-one-out testing is performed by selecting two speech files of day's folder for training and third one for testing. The resulting train and test vectors of input speech and noise estimated signal gave average classification accuracy (averaged across 26 cellphones) of around 50.8% for original-speech features (OSF) and 74.5% for noise-estimation based features (ENF).

TABLE I
CELL-PHONES USED IN THE EXPERIMENTS

Nokia	C_1(C -200),C_2 (X3-00),C_3(X2-02), C_4(5233),C_5(C2-03),C_6(C1-01), C_7(C2-00),C_8(110),C_9(X2-01), C_{10} (305),C_{11}(C1-01), C_{12}(C5-03).
Samsung	C_{13}(GT-S5360),C_{14}(GT-5801), C_{15}(GT-S3653),C_{16}(GT-S5570), C_{17}(GT-C3312),C_{18}(GTE2232), C_{19}(S3500i),C_{20}(GT-E2550), C_{21}(GT-C3222),C_{22}(GT-E2252), C_{23}(GT-E2550).
Blackberry	C_{24}(8520)
Sony	C_{25}(W150i)
Zen	C_{26} (E83-FMRM-346)

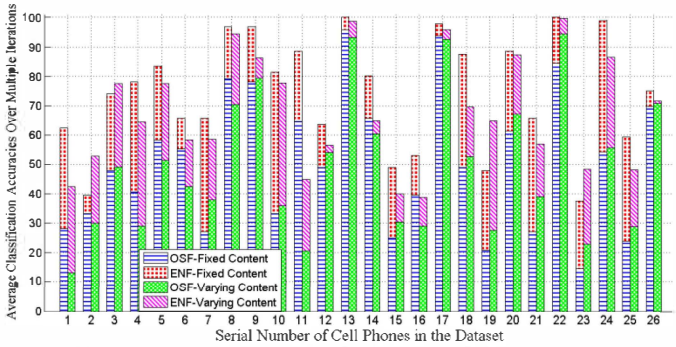

Fig. 3. Average classification accuracies for classifying all twenty-six cell phones.

This shows the efficiency of the proposed step of estimating the noise spectrum first, instead of directly working on the recorded speech. Fig. 3 shows bar graphs with the average classification accuracies for different classes for simultaneously classifying all the twenty-six cell phones. The blue bars show the accuracy obtained by original-speech features (OSF), while red bars show the accuracy obtained by noise-estimation based features (ENF). From Fig. 3, it is clear that though the improvements vary from phone to phone, the proposed step of estimating the noise spectrum improves the classification accuracy for all the cell phones (the red bars are always larger than the corresponding blue bars).

C. Classification on Varying Speech Content

The second experiment is aimed to evaluate the performance of the proposed system on recorded files with varying speech content (different spoken words). This data selection sets takes into consideration a more practical scenario where recorded speech is of varying speech content. Leave-one-out testing is performed by selecting two files from each of days, digits and months to be used for training and rest of the files are used for testing. This selection of mutually exclusive training and testing sets is repeated multiple times and average of all these cases is reported as the final average accuracies. These datasets are used to build both the models, one for direct-speech features and another for noise-estimation based features. Across twenty six phones, average classification accuracy of 49.2% and 67.7% is obtained for direct speech features and noise-estimation based features, respectively.

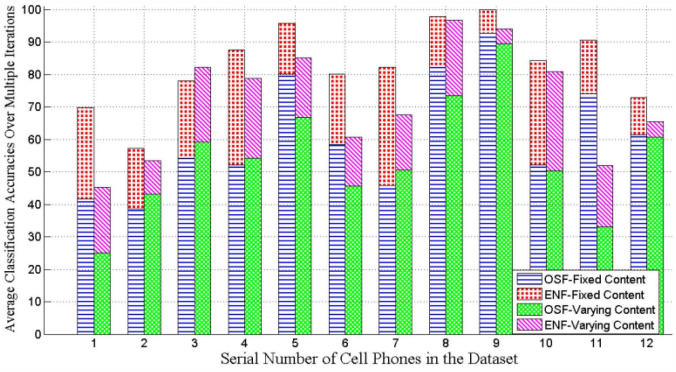

Fig. 4. Average classification accuracies for classifying twelve cell phones of same manufacturer (Nokia).

The results for this case are plotted in the form of bar graph shown in Fig. 3, showing the average classification accuracies for different classes for simultaneously classifying all the twenty-six cell phones. The green bars show the accuracy obtained by original-speech features (OSF), while magenta bars show the accuracy obtained by noise-estimation based features (ENF). As in the previous case of fixed speech content though the improvements vary from phone to phone, the proposed step of estimating the noise spectrum improves the classification accuracy for all the cell phones (the magenta bars are always larger than the corresponding green bars). Further, Fig. 3 also shows that for almost all the cell phones, there is slight decrease in the classification accuracy in case of varying speech content, in comparison to fixed speech content case. This decrease is smaller for the noise estimation based proposed system and provides a more generalizable classifier.

To check validity of the proposed system when all the cell phones belong to the same manufacturer, we performed classification on 12 cell phones from Nokia. Fig. 4 shows the average classification accuracies for different cell phones for different speech cases. The notations and trends in Fig. 4 are similar to Fig. 3, thus showing the robustness of the proposed system. For classifying 12 Nokia cell phones average classification accuracies of 54.3% and 71.9% are obtained by original speech based features and noise based features respectively. Further, the effect of variation in the manufacturer is analyzed by performing classification over a smaller subset of five cell phones from five different manufacturers. Table III shows the confusion matrix for classifying phones of different manufacturers using features from original speech signal. For classifying audio files with varying speech content, an average classification accuracy of 82.6% is obtained in this case. Table 4 shows the confusion matrix for similar experiment using features of estimated noise. For classifying audio files with varying speech content, an average classification accuracy of 90% is obtained in this case. Again, the proposed modification of using noise estimates instead of original speech for feature extraction, improves the classification accuracy.

V. Conclusions and Future Work

User-generated audio/video recordings from cell phones can be very helpful in a number of forensic applications such as securing the information left behind at a crime scene. This paper presented a system for cell-phone identification from audio recordings using MFCC features of noise estimates corresponding to each recording.

TABLE II
Confusion Matrix of Classifying Cell Phones from Different Manufacturers using Original Speech based Features

	C_1	C_{13}	C_{24}	C_{25}	C_{26}
C_1	80.8	0	4	13.4	1.9
C_{13}	0	99.3	0.7	0	0
C_{24}	4.9	0.3	73.6	5	16.2
C_{25}	14.5	0.4	3.5	74.5	7
C_{26}	3.4	1.2	6.4	4.1	84.8

TABLE III
Confusion Matrix of Classifying Cell Phones from Different Manufacturers using Features based on Estimated Noise

	C_1	C_{13}	C_{24}	C_{25}	C_{26}
C_1	93.1	0	0.7	5.7	0.5
C_{13}	0	99.9	0	0	0.1
C_{24}	1.5	0	92.3	5.6	0.7
C_{25}	4.9	1	6.8	82.9	4.3
C_{26}	4.2	3.1	6.4	4.6	81.8

The proposed system gives promising results with an average classification accuracy of around 90% for classifying cell phones belonging to different manufacturers when speech content of recorded files varies. The average classification accuracy of 72% is obtained for classifying 12 cell phones of the same manufacturer (Nokia). Future work will include different methods to predict cellphone's transfer function more precisely. Noise spectrum estimation may also be further improved using different speech enhancement methods. In addition, other methods for feature extraction apart from MFCC should be evaluated for better performance at lower computation cost.

References

[1] Lawrence Rabiner and Biing-Hwang Juang, "*Fundamentals of speech recognition*," Published by PTR prentice-Hall, Inc, New Jersey, 1993.

[2] R. O. Duda, P. E. Hart, and D. G. Stork, *"Pattern Classification*" (2nd Edition), Wiley-Interscience, 2000.

[3] Christian Kraetzer, Andrea Oermann, Jana Dittmann and Andreas Langet, "Digital Audio Forensics: a First Practical Evaluation on Microphone and Environment Classification", *Proceedings of ACM workshop on Multimedia and Security (MM & Sec),* pp. 63-74, 2007.

[4] R. Buchholz, C. Kraetzer, J. Dittmann, "Microphone classification using Fourier coefficients", *Proceedings of the 11th Information Hiding Workshop*, Vol. 5806, pp. 235-246, 2009.

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten; "*The WEKA Data Mining Software: An Update*"; SIGKDD Explorations, Volume 11, Issue 1, 2009.

[6] Christian Kraetzer, Maik Schott, Jana Dittmann, "Un-weighted Fusion in Microphone Forensics using a Decision Tree and Linear Logistic Regression Models", *Proceedings of the 11th ACM workshop on Multimedia and security (MM & Sec),* Pages 49-56, 2009.

[7] D. G. Romero and C. Y. Espy-Wilson, "Automatic acquisition device identification from speech recordings", *Proceedings IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 1806–1809, Mar. 2010.

[8] Cemal Hanilçi, Figen Ertaş, Tuncay Ertaş, and Ömer Eskidere, "Recognition of Brand and Models of Cell-phones from Recorded Speech Signals", *IEEE Transaction on Information Forensic and Security*, vol. 7, No. 2, 625-634, April 2012.

[9] Yannis Panagakis, Constantine Kotropoulos, "Automatic Telephone Handset Identification by Sparse Representation of Random Spectral Features," *Proceedings of the on Multimedia and Security (MM & Sec),* pages 91-96, September 6-7, 2012.

[10] Omer Eskidere, "Source Microphone Identification from speech recording based on Gaussian mixture model", *Turkish Journal of Electrical Engineering & Computer Sciences*, pages 1-14, Dec 2012.

[11] Y. Panagakis and C. Kotropoulos, "Telephone Handset Identification by Feature Selection and Sparse Representations", *IEEE Workshop on information forensics and security (WIFS)*, pages 73-78, Dec 2-5, 2012.

[12] S. Kamath, and P. Loizou , "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, FL, USA, Vol. 4, pp. 4164-4164, May 2002.

[13] Zeranoe, *FFmpeg*, GPL 3.0. [Online]. Available: http://ffmpeg.zeranoe.com/builds/