Mobile Phone Identification Using Recorded Speech Signals

Constantine Kotropoulos, Stamatios Samaras

Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki 54124, Greece
Email: costas@aiia.csd.auth.gr, stamatis@csd.auth.gr

***Abstract*—In this paper, we elaborate on mobile phone identification from recorded speech signals. The goal is to extract intrinsic traces related to the mobile phone used to record a speech signal. Mel frequency cepstral coefficients (MFCCs) are extracted from any recorded speech signal at a frame level. The sequences of the MFCC vectors extracted from each recording device train a Gaussian Mixture Model with diagonal covariance matrices. A Gaussian supervector is derived by concatenating the mean vectors and the main diagonals of the covariance matrices that is used as a template for each device. Experiments were conducted on a database of 21 mobile phones of various models from 7 different brands. The aforementioned database, that is called MOBIPHONE, was collected by recording 10 utterances, uttered by 12 male speakers and another 12 female speakers, randomly chosen from the TIMIT database. Three commonly used classifiers were employed, such as Support Vector Machines with different kernels, a Radial Basis Functions neural network, and a Multi-Layer Perceptron. The best identification accuracy (97.6%) was obtained by the Radial Basis Functions neural network.**

***Index Terms*—Digital speech forensics, Gaussian supervectors, Support vector machines, Radial basis functions neural network, Multi-layer perceptron**

I. Introduction

Speech is the most natural way to communicate between humans. Nowadays, low cost and sophisticated mobile phones are widespread in the society, being an indispensable communication apparatus. Mobile phones receive, transmit, store, and process information in digital form. This means that there will be lots of evidence in the speech signals recorded by mobile phones. A valuable step in digital speech forensics is phone identification, reviewed next.

First of all, one needs to extract forensic evidence about the mechanism involved in the generation of the speech recording by analyzing the speech signal [1]. That is, to identify the acquisition device by assuming that the device along with its associated signal processing chain leaves behind *intrinsic traces* in the speech signal. Indeed, the various devices (e.g., telephone handsets, mobile phones) do not have exactly the same frequency response due to the tolerance in the nominal values of the electronic components and the different designs employed by the various manufacturers [2]. This implies that the recorded speech can be considered as a signal whose spectrum is the product of the genuine speech spectrum, driving the acquisition device, and the frequency response of the latter. Consequently, the recorded speech signal can be exploited in device identification, following a blind-passive approach, as opposed to active embedding of watermarks or having access to input-output pairs [1].

Although there is a long way for making acceptable the audio forensics in the court and in that respect audio forensics are lacking behind the image forensics [3], the research on audio forensics has blossomed the last years. Several problems have attracted the interest of the forensics community, including codec identification, authentication of speakers' environment, identification of the device power source (i.e, electric network frequency (ENF)), identification of the network traversed, and automatic acquisition device identification, so far. Many studies were performed for the identification of codecs, such as MP3 [4], Windows Media Audio codec [5], Code Excited Linear Prediction codecs [6], or G.711, G.726, G.728, G.729, Internet Low-Bit codec, Adaptive Multi-Rate NarrowBand, and Silk [7]. Classification and regression trees were reported to achieve an identification accuracy of 92% among nine codecs using a 50% cross-validation on a database with 180 test conditions, comprising three noise types (car, babble and hum) at five signal to noise ratios [8]. The authentication of speakers' environment was investigated in [9]–[12]. The effectiveness of Hidden Markov Model-based phone recognition for forensic voice comparison has been evaluated in terms of both validity (accuracy) and reliability (precision) in [13]. Acoustic environment identification finds many applications (e.g., audio recording integrity authentication, real-time crime localization/identification). Statistical techniques for estimating the reverberation and background noise were proposed in [14], [15]. ENF-based techniques offer high accuracy, but they suffer from the fact that the ENF signal cannot be always extracted reliably at a frame level [16], [17]. The identification of the call origin determines whether the call traversed a cellular, a VoIP, or a PSTN network [18].

Telephone handset identification was first treated in order to avoid performance degradation in speaker recognition due to mismatches between training and test data. For example, autoassociative neural networks were reported to achieve an accuracy of 85% in a two-class problem (i.e., carbon-button vs. electret telephone handset identification) in the NIST-99 speaker evaluation database, employing 1448 test utterances [19]. A Gaussian mixture model-based handset selector was proposed in [20] and then handset-specific stochastic second-order feature transformations were applied to the distorted feature vectors increasing speaker verification accuracy. Another method for the classification of 4 microphones was originally proposed in [10] and further improved thanks to a proper fusion strategy [11]. The speech signal was parameterized by employing time domain features and the mel-frequency cepstral coefficients (MFCCs) [21]. The identification of the microphones was performed by the Naive Bayes classifier at a short-time frame level. Accuracies in the order of 60–75% were reported. Rank level fusion was shown to increase the classification accuracy to 100% [11]. The identification of 8 landline telephone handsets and 8 microphones was addressed in [1]. In particular, the intrinsic characteristics of the device were captured by concatenating the mean vectors of a Gaussian mixture model (GMM) trained on the speech recordings of each device. Linear- and mel-scaled cepstral coefficients were employed for speech signal representation. A classification accuracy of 93.2% was reported for 8 landline telephone handset identification in the Lincoln-Labs Handset Database (LLHDB) [22], when a support vector machine (SVM) classifier and a 2-fold cross-validation was employed. The identification of 14 mobile phones was proposed in [2] extracting the MFCCs from each device speech recordings, which were then classified by an SVM. An identification accuracy of 96.42% was reported for 14 different mobile phones using a set of 3360 utterances uttered by 24 speakers equally divided into a training and test set. Blind-passive methods for landline telephone handset identification were proposed in [23] and [24]. More specifically, the random spectral features were extracted by reducing the size of average log-spectrograms thanks to an orthogonal random Gaussian projection matrix [23]. In a supervised setting, the label information (i.e., the class where each device belongs to) of the training speech recordings was taken into account in order to derive a mapping between the feature space where the average log-spectrograms lie onto and the label space [24]. This supervised method reached an accuracy of 97.58% in the LLHDB. The blind-passive method for landline telephone handset identification introduced in [23] was extended by investigating the *sketches of spectral features* (SSFs) as intrinsic traces suitable for device identification in [25].

Here, the identification of mobile phones of various brands and models from recorded speech is addressed. Brand refers to the manufacturer of a mobile phone, e.g., LG, Samsung, etc. The term model refers to the product series within a brand, e.g., LG L9, Samsung Galaxy Nexus S, and so on. To do so, intrinsic traces related to the device used in speech signal acquisition should be derived by modeling the MFCCs that are extracted from any recorded speech signal at a frame level. The MFCCs have been dominantly used in speech recognition, speaker recognition, and acquisition device identification despite the fact that the aforementioned tasks seek different types of information. They encode the frequency content of the signal by parameterizing the rough shape of its spectral envelope. Starting with the short-term power spectrum of the speech signal, discrete cosine transform is applied to the log power spectrum at the output of a filter bank in a nonlinear mel-warped frequency scale. We resort to GMMs with diagonal covariance matrices in order to model the probability density function of the MFCC vectors. Having training a GMM for each device, a Gaussian supervector (GSV) is built by concatenating the mean vectors and the main diagonals of the covariance matrices of all components. The GSVs are extracted without resorting to a GMM universal background model [26] and are used as recording device templates. A database of 21 mobile phones of various models from 7 different brands was collected by recording 10 utterances, uttered by 12 male speakers and another 12 female speakers, randomly chosen from the TIMIT database [27]. Let us call this database MOBIPHONE, hereafter. Experiments were conducted on the MOBIPHONE by employing three commonly used classifiers, such as SVMs with different kernels, a Radial Basis Functions neural network (RBF-NN), and a Multi-Layer Perceptron (MLP). The top identification accuracy of 97.6% was achieved by the RBF-NN.

The main contribution of the paper is in the disclosure of experimental evidence for mobile phone identification, employing the aforementioned classifier and the release of the MOBIPHONE database that is made publicly available. The rest of the paper is organized as follows. In Section 2, an overview of the identification system is presented. Feature extraction is detailed in Section 3. The MOBIPHONE database is described in Section 4. Experimental results are disclosed in Section 5 and conclusions are drawn in Section 6.

II. Mobile Phone Identification

A mobile phone identification system consists of three modules: feature extraction, feature modeling, and the classifier. Feature extraction is the process of extracting information related to the acquisition device that the recorded speech signals bear. Here, this information is captured by the MFCCs. Modeling aims at deriving a parametric model for the probability density function of the MFFC vectors, i.e., a GMM. The model parameters are the mixture weights, the mean vectors, and the diagonal covariance matrices. The mean vectors and the main diagonals of the covariance matrices are exploited to build the GSVs. The GSVs are split into a training and a test set. The training set is used to train the classifier, while the test set is used to assess the classifier performance. The entire processing chain is depicted in Figure 1. Feature modeling and classifier are referred collectively as similarity measure.

III. Feature Extraction

The underlying hypothesis is that the mobile phones leave behind intrinsic traces in the speech signal. These traces can be modeled and detected by pattern recognition techniques [1]. As said previously, the MFCCs are extracted from recorded speech signals. Let us elaborate on the appropriateness of the MFCCs for acquisition device characterization. Assume

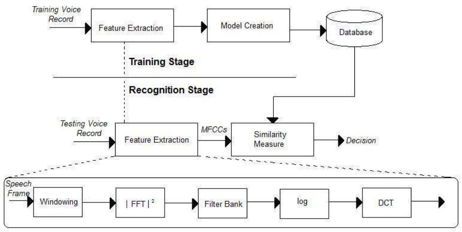

Fig. 1. Mobile phone identification recognition system.

that a mobile phone is a linear time-invariant system with impulse response $h[n]$. If $x[n]$ is the speech signal uttered by a speaker, the recorded speech signal by the device $y(n)$ is the convolution of $x[n]$ with the impulse response, i.e.,

$$y[n] = (h * x)[n]. \tag{1}$$

Because the speech is not a stationary signal, it is divided into overlapped segments of duration 20 ms with a hop size of 10 ms, known as frames. The speech frames are obtained by multiplying the speech signal with a Hamming window. Then the pth frame of the recorded speech signal is given by

$$y_p[n] = (x[n]\, w[pN - n]) * h[n] \tag{2}$$

where the term inside parentheses is identified as the pth frame of the speech signal and $w[pN - n]$ denotes the window ending at the sample pN, N being the window length. Since the identity of the recording mobile phone is embedded into the recorded signal through a convolution, cepstrum looks appropriate to separate the intrinsic trace left in the recorded signal by the acquisition device [28]. Taking the discrete-time Fourier transform of both sides of (2), we obtain

$$Y_p(f) = \mathcal{F}\left\{x[n]\, w[pN - n]\right\}\; H(f) \tag{3}$$

where $\mathcal{F}\{\cdot\}$ denotes the discrete-time Fourier transform, $Y_p(f) = \mathcal{F}\{y_p[n]\}$, and $H(f) = \mathcal{F}\{h[n]\}$ is the frequency response of the mobile phone. (3) can be written as

$$Y_p(f) = \left[\int_0^1 X(\theta)\, W_p(\theta - f)\, d\theta\right] H(f) \tag{4}$$

where $W_p(f) = \mathcal{F}\left\{w[pN - n]\right\}$. Using the properties of the discrete-time Fourier transform

$$w[-n] \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad W^*(-f) \tag{5}$$

$$w[n - n_0] \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad e^{-j\,2\pi\, f n_0}\, W(f) \tag{6}$$

and substituting into (4) we arrive at

$$Y_p(f) = \left[\int_0^1 X(\theta) W^*(\theta - f)\, e^{j2\pi(\theta - f)pN}\, d\theta\right] H(f). \tag{7}$$

Let us denote the integral in (7) by $\tilde{X}_p(f)$. The discrete-time Fourier transform $X(f)$ is expressed as the product of the discrete-Fourier transform of the excitation signal $E(f)$ and the frequency response of the vocal tract $V(f)$. By invoking the arguments in [28], $\tilde{X}_p(f)$ can be approximated as

$$\tilde{X}_p(f) = \int_0^1 E(\theta)V(\theta)W^*(\theta - f)e^{j2\pi(\theta - f)pN} d\theta \approx E_p(f)V(f) \tag{8}$$

where $E_p(f) = \mathcal{F}\{e[n]w[pN - n]\}$ denotes the discrete-time Fourier transform of the pth frame of the excitation signal. The substitution of (8) into (7) yields

$$Y_p(f) = E_p(f)\, V(f)\, H(f). \tag{9}$$

That is, it has been proved that the recording by a mobile phone leaves behind an intrinsic trace in the recorded speech spectrum.

Proceeding to the mel cepstrum, it is interesting to note that the MFCCs have been the baseline features for speech recognition or speaker verification (i.e., deconvolution of $V(f)$ in (9)), speech emotion recognition (i.e., deconvolution of $E_p(f)$ in (9)), and device acquisition (i.e., deconvolution of $H(f)$ in (9)). The MFCC calculation employs frames of duration 20 ms with a hop size of 10 ms, and a 42-band filter bank. The correlation between the frequency bands is reduced by applying the discrete cosine transform to the log-energies of the bands. Sequences of 23-dimensional MFCCs are extracted per frame. The histogram of the 5th MFCC for different mobile phones and the same recorded speech signal is plotted in Figure 2. It is self-evident that the histograms differ across the various mobile phone brands and models. Having extracted a sequence of 23-dimensional MFCC vectors for all frames of recorded speech utterances by each mobile phone, a GMM was trained by means of the Expectation-Maximization algorithm [29].

IV. MOBIPHONE DATABASE

The MOBIPHONE database contains 21 mobile phones of various models from 7 different brands. The brands and models of the mobile phones are listed in Table I. The 7 different brands include some of the major companies in the market of mobile communications, like Samsung, Nokia, LG, and Apple. Accordingly, the MOBIPHONE is a representative sample of the mobile phone industry worldwide.

For 12 male speakers and another twelve female speakers, randomly chosen from the TIMIT database [27], whose identities are listed in Table II, 10 utterances were recorded by the various mobile phones. The recordings were made in a silent controlled environment with the same recording equipment. Each speaker reads 10 sentences approximately of 3s long. The first two sentences are the same for every speaker, but the rest 8 are different. The raw recordings were in adapted multi-rate (AMR) format and were later converted into Waveform Audio File (WAV) format. The sampling frequency was 16 kHz. In the first release of the MOBIPHONE[1], the 10 utterances per speaker were concatenated in a single 30s long recording, yielding 504 recordings all together. In the second release

[1] https://www.dropbox.com/sh/9n7fy7moi825bgk/WFLBKxUitV

TABLE I
BRANDS AND MODELS OF THE MOBILE PHONES IN THE MOBIPHONE AND THEIR CLASS NAMES.

Class Name	Brand and Model	Class Name	Brand and Model
HTC1	HTC desire c	APPLE1	iPhone5
HTC2	HTC sensation xe	S1	Samsung E2121B
LG1	LG GS290	S2	Samsung E2600
LG2	LG L3	S3	Samsung GT-I8190 mini
LG3	LG Optimus L5	S4	Samsung GT-N7100 (Galaxy Note2)
LG4	LG Optimus L9	S5	Samsung Galaxy GT-I9100 s2
N1	Nokia 5530	S6	Samsung Galaxy Nexus S
N2	Nokia C5	S7	Samsung e1230
N3	Nokia N70	S8	Samsung s5830i
SE1	Sony Ericsson c902	V1	Vodafone joy 845
SE2	Sony Ericsson c510i		

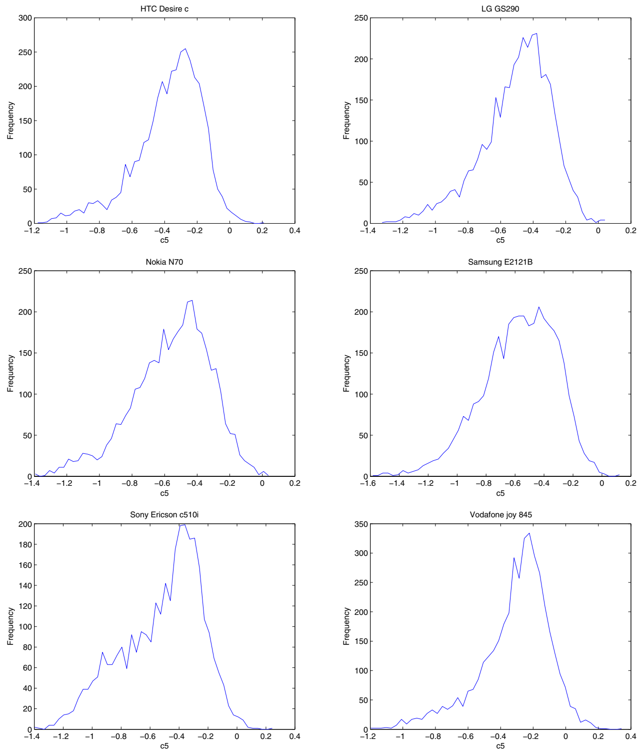

Fig. 2. Histogram of the 5th MFCC for different mobile phones and the same recorded speech signal.

of the MOBIPHONE, the recordings will be provided per utterance.

V. Experimental Evaluation

Experiments were conducted on the MOBIPHONE by employing three commonly used classifiers, such as SVMs with different kernels [30], [31], an RBF-NN [32], and an MLP [32]. Two disjoint subsets of 252 recordings were created that are balanced in the number of recordings as well as speakers and gender. The first subset was used during the training phase to train the classifiers, while the second one was used to assess classifier performance in 252 claims.

TABLE II
TIMIT SPEAKER IDENTITIES.

Speaker	TIMIT ID	Speaker	TIMIT ID
1	TEST\DR1\FAKS0	13	TRAIN\DR4\MJAC0
2	TEST\DR2\MPDF0	14	TRAIN\DR1\FVMH0
3	TRAIN\DR1\MMGG0	15	TRAIN\DR1\FETB0
4	TRAIN\DR1\MMRP0	16	TRAIN\DR1\FKFB0
5	TRAIN\DR2\MDMT0	17	TRAIN\DR2\FAEM0
6	TRAIN\DR2\MKAJ0	18	TRAIN\DR2\FCYL0
7	TRAIN\DR2\MRJM1	19	TRAIN\DR3\FALK0
8	TRAIN\DR3\MLNS0	20	TRAIN\DR3\FCKE0
9	TRAIN\DR3\MREH1	21	TRAIN\DR3\FDFB0
10	TRAIN\DR1\MRWS0	22	TRAIN\DR3\FSJS0
11	TRAIN\DR4\MSMS0	23	TRAIN\DR4\FCAG0
12	TRAIN\DR4\MAEB0	24	TEST\DR8\FJSJ0

To begin with, a multiclass SVM was trained and tested for closed set identification. Linear, RBF with $\sigma = 5$, and third order polynomial kernels were used in SVMs for various GMM mixture components. Two types of GSVs were tested, namely GSVs formed by concatenating the mean vectors of the GMM components and GSVs including both the mean vectors and the main diagonal of the covariance matrices. In the former case, Table III summarizes the accuracies measured, while Table IV lists the accuracies achieved in the latter case. The choices for the kernels, the GMM components, and the type of GSVs influence both the identification accuracy and the time needed to perform identification. In particular, the fewer GMM components, the higher accuracy is measured. The RBF kernel yields the top accuracy for both GSV types. Including information from the covariance matrices in the GSV improves slightly the accuracy. The top accuracy achieved with SVMs was 92.5%.

For RBF-NN, a network of 21 hidden neurons was employed to perform closed set identification. The identification accuracies for different values of σ and GSV types are presented in Tables V and VI. All the three factors affect the identification accuracy. It is seen, that just one Gaussian component with a GSV formed by concatenating the mean vector and the variances and a small value for the spread of

TABLE III
IDENTIFICATION ACCURACIES (IN %) ACHIEVED BY THE SVM FOR GSVS FORMED BY MEAN VECTORS ONLY.

GMM Components	Kernel function type	Accuracy
1	Linear	84.1
3	Linear	79.8
6	Linear	75.3
1	RBF	*92.1*
3	RBF	61.9
6	RBF	21.4
1	Polynomial	*92.1*
3	Polynomial	74.6
6	Polynomial	55.2

TABLE IV
IDENTIFICATION ACCURACIES (IN %) ACHIEVED BY THE SVM FOR GSVS INCLUDING ALSO THE MAIN DIAGONALS OF THE COVARIANCE MATRICES.

GMM Components	Kernel function type	Accuracy
1	Linear	90.4
3	Linear	77.3
6	Linear	78.9
1	RBF	**92.5**
3	RBF	79.3
6	RBF	7.1
1	Polynomial	89.2
3	Polynomial	59.1
6	Polynomial	34.3

RBFs yields the top accuracy. RBF-NN was found to be the best classifier, achieving an accuracy of 97.6%.

TABLE V
IDENTIFICATION ACCURACIES (IN %) ACHIEVED BY THE RBF-NN FOR GSVS FORMED BY MEAN VECTORS ONLY.

GMM Components	σ	Accuracy
1	0.1	*97.2*
3	3	81.7
6	4	85.5

Finally, an MLP was tested for closed set identification. The best neural network parameter settings are disclosed in Table VII. MLP performance is influenced by the proper choice of the 5 factors included in Table VII. That is, the type of GSV, the momentum and learning rate in the backpropagation algorithm, the number of hidden neurons, and number of the epochs. Accordingly, the tuning of MLP is more cumbersome than that of the other classifiers, requiring multiple tests. MLP is found to be the second best classifier achieving a top accuracy of 96.4%.

TABLE VI
IDENTIFICATION ACCURACIES (IN %) ACHIEVED BY THE RBF-NN FOR GSVS INCLUDING ALSO THE VARIANCES.

GMM Components	σ	Accuracy
1	0.1	**97.6**
3	5	73.8
6	5	74.2

In order to check if the top accuracy differences are statistically significant, the approximate analysis in [33] is applied. Assume that the accuracies ϖ_1 and ϖ_2 are binomially distributed random variables. If $\hat{\varpi}_1, \hat{\varpi}_2$ denote the empirical accuracies, and $\overline{\varpi} = \frac{\hat{\varpi}_1+\hat{\varpi}_2}{2}$, the hypothesis $H_0 : \varpi_1 = \varpi_2 = \overline{\varpi}$ is tested at 95% level of significance. The accuracy difference has variance $\beta = 2\frac{\overline{\varpi}(1-\overline{\varpi})}{M}$, where M is the number of test recordings (i.e., 252). For $\varphi = 1.65\sqrt{\beta}$, if $\hat{\varpi}_1 - \hat{\varpi}_2 \geq \varphi$, we reject H_0 with risk 5% of being wrong. The aforementioned analysis yields that the performance gain between the RBF-NN and the SVM ($\varphi = 3.19\%$) as well as between the MLP and the SVM ($\varphi = 3.36\%$) are statistically significant. On the contrary, the performance gain between the RBF-NN and the MLP ($\varphi = 2.5\%$) is not statistically significant.

VI. CONCLUSIONS

A publicly available database, the MOBIPHONE database, that contains 21 mobile phones of various models from 7 different brands has been released. Very promising mobile phone identification accuracy has been obtained by three commonly used classifiers, namely the RBF-NN, the MLP, and the SVMs with different kernels. The top accuracy of 97.6% has been achieved by the RBF-NN. The performance gain between the RBF-NN and the SVM as well as between the MLP and the SVM has been attested to be statistically significant. Future research will address the impact in identification accuracy of factors, such as unknown background noise, gain settings in the mobile phone handset, or the speech coding algorithm, deviating from the ideal situation studied here and converging to a real forensic scenario.

Acknowledgments. This work has been supported by the Cost Action IC 1106 “Integrating Biometrics and Forensics for the Digital Age”.

REFERENCES

[1] D. Garcia-Romero and C. Y. Espy-Wilson, “Automatic acquisition device identification from speech recordings,” in *Proc. 2010 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Dallas, TX, USA, 2010, pp. 1806–1809.

[2] C. Hanilci, F. Ertas, T. Ertas, and O. Eskidere, “Recognition of brand and models of cell-phones from recorded speech signals,” *IEEE Trans. Information Forensics and Security*, vol. 7, no. 2, pp. 625–634, 2012.

[3] R. Maher, “Audio forensic examination,” *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 84–94, 2009.

TABLE VII
IDENTIFICATION ACCURACIES (IN %) ACHIEVED BY THE MLP FOR BOTH TYPES OF GSVS (M STANDS FOR THE GSVS FORMED BY THE MEAN VECTORS ONLY AND MC STANDS FOR THE GSVS, WHICH INCLUDE THE VARIANCES AS WELL).

GMM Components	GSV Type	Momentum	Learning Rate	Hidden Neurons	Epochs	Accuracy
1	M	0.9	0.1	150	250	96
2	MC	0.9	0.1	150	250	**96.4**

[4] R. Yang, Z. Qu, and J. Huang, "Detecting digital audio forgeries by checking frame offsets," in *Proc. 10th ACM Multimedia and Security Workshop*, New York, NY, USA, 2008, pp. 21–26.

[5] D. Luo, W. Luo, R. Yang, and J. Huang, "Compression history identification for digital audio signal," in *Proc. 2012 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Kyoto, Japan, 2012, pp. 1733–1736.

[6] J. Zhou, D. Garcia-Romero, and C. Y. Espy-Wilson, "Automatic speech codec identification with applications to tampering detection of speech recordings," in *Proc. 12th INTERSPEECH*, Florence, Italy, 2011, pp. 2533–2536.

[7] F. Jenner and A. Kwasinski, "Highly accurate non-intrusive speech forensics for codec identifications from observed decoded signals," in *Proc. 2012 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Kyoto, Japan, 2012, pp. 1737–1740.

[8] D. Sharma, P. A. Naylor, N. D. Gaubitch, and M. Brookes, "Non intrusive codec identification algorithm," in *Proc. 2012 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Kyoto, Japan, 2012, pp. 4477–4480.

[9] A. Oermann, A. Lang, and J. Dittmann, "Verifier-tuple for audio-forensic to determine speaker environment," in *Proc. 7th ACM Multimedia and Security Workshop*, New York, NY, USA, 2005, pp. 57–62.

[10] C. Kraetzer, A. Oermann, J. Dittmann, and A. Lang, "Digital audio forensics: a first practical evaluation on microphone and environment classification," in *Proc. 9th ACM Multimedia and Security Workshop*, Dallas, TX, USA, 2007, pp. 63–74.

[11] C. Kraetzer, M. Schott, and J. Dittmann, "Unweighted fusion in microphone forensics using a decision tree and linear logistic regression models," in *Proc. 11th ACM Multimedia and Security Workshop*, Princeton, NJ, USA, 2009, pp. 49–56.

[12] H. Malik and H. Farid, "Audio forensics from acoustic reverberation," in *Proc. 2010 IEEE Int. Conf. Acoustics Speech and Signal Processing*, Dallas, TX, USA, 2010, pp. 1710–1713.

[13] C. C. Huang and J. Epps, "A study of automatic phonetic segmentation for forensic voice comparison," in *Proc. 2012 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Kyoto, Japan, 2012, pp. 1853–1856.

[14] H. Zhao and H. Malik, "Acoustic recording location identification using acoustic environment signature," *IEEE Trans. Information Forensics and Security*, vol. 8, no. 11, pp. 1746–1759, 2013.

[15] H. Malik, "Acoustic environment identification and its applications to audio forensics," *IEEE Trans. Information Forensics and Security*, vol. 8, no. 11, pp. 1827–1837, 2013.

[16] A. J. Cooper, "Further considerations for the analysis of ENF data for forensic audio and video applications," *The Int. J. of Speech, Language, and the Law*, vol. 18, no. 1, pp. 99–120, 2011.

[17] O. Ojowu, Jr., J. Karlsson, J. Li, and Y. Liu, "ENF extraction from digital recordings using adaptive techniques and frequency tracking," *IEEE Trans. Information Forensics and Security*, vol. 7, no. 4, pp. 1330–1338, 2012.

[18] V. A. Balasubramaniyan, P. Amit, M. Ahamad, M. Hunter, and P. Tranyo, "PinDr0p: Using single-ended audio features to determine call provenances," in *Proc. 17th ACM Conf. Computer Communications*, Chicago, IL, 2010, pp. 109–120.

[19] S. Kishore and Y. B., "Identification of handset type using autoassociative neural networks," in *Proc. 4th Int. Conf. Advances in Pattern Recognition and Digital Techniques*, 1999, pp. 353–356.

[20] M.-W. Mak and S.-Y. Kung, "Combining stochastic feature transformation and handset identification for telephone-based speaker verification," in *Proc. 2002 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. I, Orlando, FL, 2002, pp. 701–704.

[21] S. B. Davies and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. ASSP*, vol. 28, no. 4, pp. 357–366, 1980.

[22] D. Reynolds, "HTIMIT and LLHDB: speech corpora for the study of handset transducer effects," in *Proc. 1997 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, Munich, Germany, 1997, pp. 1535–1538.

[23] Y. Panagakis and C. Kotropoulos, "Automatic telephone handset identification by sparse representation of random spectral features," in *Proc. 14th ACM Multimedia and Security Workshop*, Coventry, U.K., 2012, pp. 91–95.

[24] ——, "Telephone handset identification by feature selection and sparse representations," in *Proc. 2012 IEEE Int. Workshop Information Forensics and Security*, Tenerife, Spain, 2012, pp. 73–78.

[25] C. Kotropoulos, "Source phone identification using sketches of features," *IET Biometrics*, 2014.

[26] D. A. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[27] J. Garofolo, "Getting started with the DARPA TIMIT cd-rom: An acoustic phonetic continuous speech database," National Inst. Standards and Technology (NIST), Tech. Rep., 1988.

[28] J. R. Deller, Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. New York, NY, USA: Wiley-Interscience-IEEE, 2000.

[29] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm (with discussion)," *Journal Royal Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.

[30] V. Vapnik, *Statistical Learning Theory*. New York, NY, USA: J. Wiley & Sons, 1998.

[31] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions Intelligent System Technologies*, vol. 2, no. 3, pp. 1–27, 2011.

[32] S. Haykin, *Neural Networks and Learning Machines, 3/e*. Upper Saddle River, N.J., USA: Prentice Hall, 2008.

[33] I. Guyon, J. Makhoul, R. Schwartz, and V. Vapnik, "What size test set gives good error rate estimates?" *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 52–64, 1998.