TÜBİTAK

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2014) 22: 754 – 767
© TÜBİTAK
doi:10.3906/elk-1207-74

Source microphone identification from speech recordings based on a Gaussian mixture model

Ömer ESKİDERE*
Department of Electrical and Electronics Engineering, Bursa Orhangazi University, Yıldırım, Bursa, Turkey

Received: 20.07.2012 • **Accepted:** 11.12.2012 • **Published Online:** 21.03.2014 • **Printed:** 18.04.2014

Abstract: Microphone identification is a specific type of media forensics that investigates whether it is possible to identify the source microphone from speech recordings. The main aim of this study is to find out which of the several feature extraction techniques are best suited to the source microphone identification systems. We perform microphone identification experiments with 16 different microphones using 3 datasets. In order to improve the results on the datasets, we also investigate the important parameters that may affect the microphone identification performance. Our experimental results show that the proposed method is comparable to the existing studies in a closed-set identification rate.

Key words: Gaussian mixture model, microphone identification, and microphone forensics

1. Introduction

Media forensics, with the development of electronic and computer science, has become a hot topic in the field of civil or criminal law enforcement investigations. Over the past few decades, the research on media forensics [1,2] has basically focused on image forensics (image authentication and source camera identification). In comparison to the studies on image forensics, the methods for authenticating audio recordings and source microphone identification are less developed.

Digital audio recordings can be used in criminal investigations such as bribery, political corruptions, drug deals, and other racketeering activities. In some cases, the collected digital audio, which could be recorded unintentionally or during a surveillance operation, may need to be evaluated for potential use as digital evidence in court. In forensics and criminal investigations, determining the integrity (of the original or the copy) and authenticity (referring to the ability to confirm the integrity of information) of a digital audio, 2 important issues arise. The first is the establishment of whether or not the audio has undergone any form of alteration or manipulation after it was initially recorded. This could be achieved by employing methods for audio authentication, such as bispectral analysis [3], electric network frequency-based approaches [4–6], and a number of active spectral coefficients as a function of the frame offset for MP3 files [7]. The second issue in audio forensics is the identification of acquisition devices. The determination of an acquisition device would particularly be useful as evidence in court to establish the source of the audio. Moreover, the traces produced by the acquisition device due to alterations and/or manipulations can help forensic examiners in trying to verify the integrity of the content. Therefore, determining the acquisition device can be a significant step in the criminal investigation of the evidence.

*Correspondence: omer.eskidere@bou.edu.tr

A microphone leaves behind some specific traces, due to its intrinsic characteristics, in all of the audios that it records. These digital traces may be due to component technologies, component imperfections, defects, and other manufacturing errors. The traces left in the speech signal by the source microphone are used in the field of speaker recognition [8,9]. The focus of the current study is to identify the source microphone using speech recordings only.

For source microphone identification, Kraetzer et al. [10] proposed a novel idea based on unsupervised and supervised learning methods. In their method, each audio was represented through a segmental feature extractor, which is normally used in steganalysis. They used time-domain features and mel-frequency cepstral coefficients (MFCCs). The k-means algorithm and the naïve Bayes classifier were applied to classify audios from 4 different microphones. The best correct microphone classification results were achieved as 75.99% by the Bayesian classification and as 41.57% by the k-means clustering. In [11], the authors showed that fast Fourier transform coefficients could be used in order to determine the microphone model. They tested 7 different microphones and achieved a 93.5% correct classification rate with linear regression models. In [12], it was pointed out that fusion operations, such as the match level, rank level, and decision level, could be implemented for reliable microphone classification. Using the same microphones that were used in [10,11], 100% accuracy was reached via the method of rank level fusion. Moreover, Garcia-Romero and Espy-Wilson [13] investigated the performance of MFCCs and linear scale cepstral coefficients with the support vector machine (SVM) classifier. They obtained classification accuracies of higher than 90% for landline telephone handsets and for 8 different microphones. In addition to their previous studies, Kraetzer et al. [14] presented a context model, for microphone forensics, in which they explored the identification and determination of suitable classification methods and features.

We recently addressed a new problem of recognizing cell phones from recorded speech signals [15]. Vector quantization and SVM-based classification algorithms are used in several experiments, and, as a result, an identification rate of 96.42% is achieved on a set of 14 models of cell phones.

Some characteristics of microphone identification have not yet been examined in the literature. Therefore, the main aim of the present study is to find out which of the several feature extraction techniques are best suited to source microphone identification systems. Although MFCCs have recently been exploited successfully in microphone forensics [10–13], our focus is on linear prediction cepstrum coefficients (LPCCs) and perceptually-based linear predictive coefficients (PLPCs), thanks to their widespread use and easy modeling. Additionally, we aim to find out the important parameters that affect the performance of microphone recognition. For this purpose, the model order of the Gaussian mixture model (GMM) is investigated. The impact of the training duration and the test utterance lengths are also tested on the microphone identification.

In section 2, the source microphone identification scheme proposed here is described, including the speech recording, feature extraction, microphone model generation, and decision. In section 3, the test procedure and the test setup are described. Section 4 presents the experimental results on speech samples recorded from various scenarios. Comparisons with existing studies and discussions and conclusions are given in sections 5 and 6, respectively.

2. Method

Source microphone identification from speech recordings is the main objective of this study, whose outline is depicted in Figure 1.

Initially, the speech wave is transformed into an analog speech signal using a microphone. It is then sampled to form a digital speech signal by an analog to digital converter (ADC). Next, feature extraction is applied for the training and test speech datasets. In the training phase, for each source microphone, a model is generated using the extracted features. Finally, the extracted features are fed into a GMM classifier to determine the originating microphone of the speech being tested. The stages of the proposed method are described below.

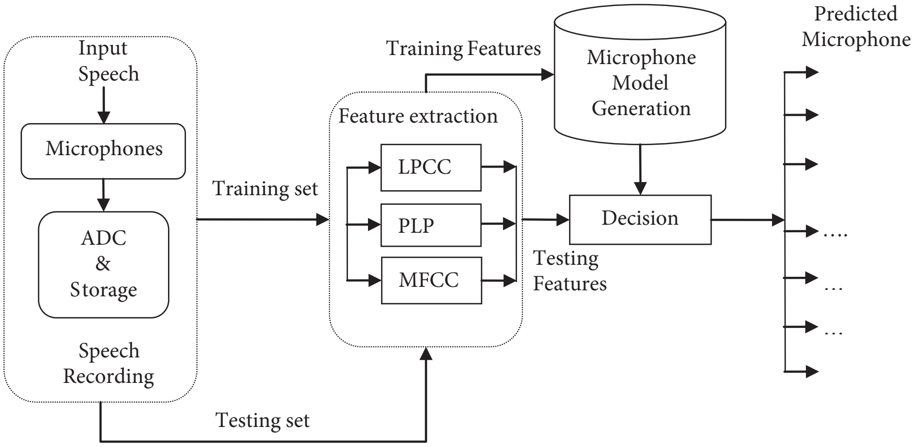

Figure 1. The work flow of our proposed method.

2.1. Speech recording

2.2. Microphones

This section covers how microphone properties affect the microphone identification system. A microphone is the most critical element of the recording stage, and it is a transducer that converts sounds into equivalent electrical output signals. Microphones have some key characteristics, which include the transducer type, sensitivity, frequency response, and directionality (polar pattern). All of these parameters significantly affect the recognition performance of the source microphones.

The first parameter examined is the transducer type. There are different microphone transducer types. The most widely used are dynamic, condenser, and electret condenser. The basic working principles of these microphones are as follows:

Sound pressure variations through the diaphragm caused the motion of the voice coil in a dynamic microphone. Thus, this motion in the magnetic field produces an electrical signal. A condenser microphone comprises a very thin diaphragm and a stationary back plate that has an external bias voltage. On the other hand, electret condenser microphones consist of an electret layer that is capable of holding a fixed electric charge [16]. Thus, microphones that have a capacitor convert the capacitance variations into voltage variations, proportional to the diaphragm input pressure.

The second parameter deals with the sensitivity, which indicates how well the microphone can convert the sound pressure into electricity. The sensitivity of a condenser microphone depends on the electrical sensitivity of the microphone and the mechanical sensitivity of the microphone diaphragm. While the electrical sensitivity is the correlation of the thickness of the air gap between the 2 plates to the bias voltage, the mechanical sensitivity is related to the deflection of the diaphragm [17–19]. Moreover, electret condenser microphones have an integrated field-effect transistor preamplifier that acts as an impedance converter and amplifies the signal. The measured sensitivity of an electret condenser microphone also affects the gain of the preamplifier [20].

Another characteristic parameter of a microphone is the frequency response, referring to the frequencies that a microphone is capable of capturing effectively. It indicates the sensitivity of a microphone to sound (e.g., pink noise) at different frequencies. Microphones have 2 main types of frequency responses: flat frequency response and tailored frequency response. An ideal flat frequency response denotes all audible frequencies (20 Hz–20 kHz) that are at the same output level. However, the tailored frequency response consists of a peak in the particular frequency range, such as 2–8 kHz.

Finally, a microphone's directionality indicates the microphone's sensitivity to sound from various directions. In general, most microphones can be placed in 1 of 3 main groups: omnidirectional, unidirectional, and bidirectional. An omnidirectional microphone picks up sound from almost every direction, whereas a unidirectional microphone picks up sound predominantly from one direction, while rejecting the sound that arrives from other directions. Finally, a bidirectional microphone picks up sound from 2 opposite directions, while side sensitivities are weaker.

2.3. Feature extraction

In this section, the features used in this study to identify the source microphone model from the recorded speech signal are described. The extraction of these features is an important task and it significantly affects the recognition performance. The widely used spectral based features that were previously employed for speech and speaker recognition problems are covered in [21–23]. These features include LPCCs, PLPCs, and MFCCs.

A LPCC is derived from linear prediction coefficients (LPCs), which is the weighted linear combination of the previous speech samples. As a first step, the speech sample is decomposed into frames and the LPC can directly be calculated from these windowed frames. Next, the LPC is converted into a LPCC using the equations defined in [21]. Different from the LPCC algorithm, the PLPC algorithm includes equal loudness preemphasis, critical-band integration, and intensity-to-loudness compression stages [22]. In the MFCC algorithm, the speech signal is divided into frames and the amplitude values of the discrete Fourier transform of the signal are calculated for each windowed frame. Next, the amplitude values are converted to mel-filter bank outputs, and the output from each filter is log-compressed and transformed via the discrete cosine transform to cepstral coefficients. The details of these feature extraction techniques may be found in [23–25].

The differences on features between the microphones (see Table 1) are visually illustrated by histograms. The histograms of the third LPCC feature (c_3) are given in Figure 2. The same speech data from a particular speaker are used to show the behavior of the same feature on each microphone.

In Figure 2, it is shown that each microphone has distinctive features. As different microphones generate various behaviors on the same feature, the results lead to different histograms.

2.4. Model description

The GMM has proved to be a powerful technique for pattern classification and data modeling [26]. For a D-dimensional feature vector $\vec{s}$, a Gaussian mixture density is given by the following equation:

$$p\left(\vec{s}\,\middle|\,\lambda\right) = \sum_{i=1}^{N} c_i d_i\left(\vec{s}\right), \tag{1}$$

where $N$is the number of mixture components, λ is the microphone class model, $c_i\epsilon\ [0,\ 1]$ is the mixture weights, and $d_i(\vec{s})$ is the component densities, which can be expressed as:

$$d_i\left(\vec{s}\right) = \frac{1}{\left(2\pi\right)^{D/2}\left|\Sigma_i\right|^{1/2}} exp\left\{-\frac{1}{2}\left(\vec{s}-\vec{\mu}_i\right)' \Sigma_{\mathrm{i}}^{-1}\left(\vec{s}-\vec{\mu}_i\right)\right\}. \tag{2}$$

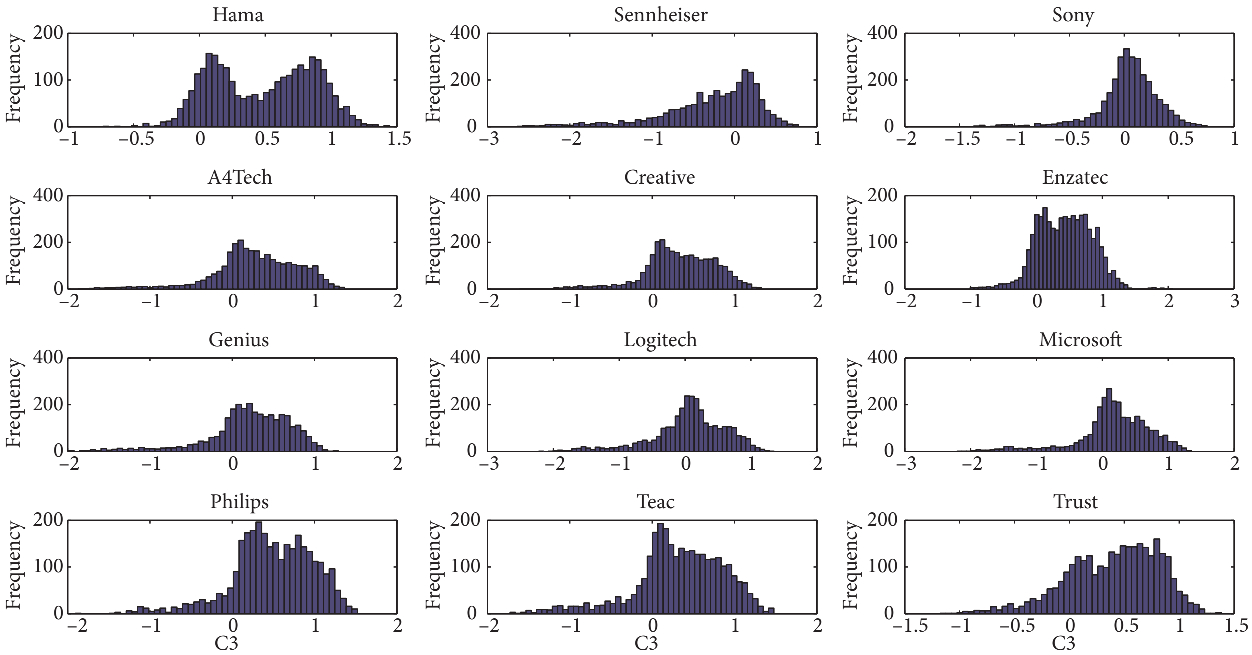

Figure 2. Histograms of the third LPCC feature for each microphone model.

Each microphone is designated by a model parameterized by the mean vectors $\vec{\mu}_i$, covariance matrices $\sum_i$, and mixture weights, which is represented by the notation:

$$\lambda = \{c_i,\, \mu_i, \Sigma_i\} \qquad i = 1,\, 2,\, \ldots,\, N \tag{3}$$

During the training phase of the GMM, model parameters (λ) are estimated from the training data via the expectation maximization (EM) iterative procedure. The EM algorithm computes the maximum likelihood, and it enables a monotonic increase in the model's log-likelihood value [26]. For a given sequence training feature, vector $S = \{\vec{s}_1, \ldots, \vec{s}_T\}$ begins with the initial model λ, estimates a new model $\bar{\lambda}_,$ and this is repeated for the following iterations. In this case, the a posteriori probability for mixture i can be defined as:

$$p\,(i_t = i\,|\,\vec{s}_t, \lambda) = \frac{c_i d_i\,(\vec{s}_t)}{\sum\limits_{k=1}^{N} c_k d_k\,(\vec{s}_t)}. \tag{4}$$

The mixture weights, $\bar{c}_i$, can be updated as:

$$\bar{c}_i = \frac{1}{T}\sum_{t=1}^{T} p\,(i_t = i\,|\,\vec{s}_t, \lambda). \tag{5}$$

The means that $\bar{\mu}_i$ can be updated as:

$$\bar{\mu}_i = \frac{\sum\limits_{t=1}^{T} p\,(i_t = i\,|\,\vec{s}_t, \lambda)\vec{s}_t}{\sum\limits_{t=1}^{T} p\,(i_t = i\,|\,\vec{s}_t, \lambda)}. \tag{6}$$

The covariance matrices, $\bar{\Sigma}_{\mathrm{i}}$, can be updated as:

$$\bar{\Sigma}_{\mathrm{i}} = \frac{\sum_{t=1}^{T} p\,(i_t = i\,|\,\vec{s}_t, \lambda)\vec{s}_t\vec{s}_t^{\,\prime}}{\sum_{t=1}^{T} p\,(i_t = i\,|\,\vec{s}_t, \lambda)} - \vec{\bar{\mu}}_i\vec{\bar{\mu}}_i^{\,\prime}. \tag{7}$$

For each iteration, Eqs. (4–7) are repeated until a convergence threshold is reached.

The decision phase of the GMM comprises the calculation of a simple set of likelihood functions using the test speech and the GMM for each microphone. In this phase, a group of microphones, $M = \{1, 2, \ldots, M\}$, is denoted by GMM's$\lambda\lambda\lambda_{12,\ldots,\mathrm{M}}$. Given a sequence test feature, vector $S = \{\vec{s}_1, \ldots, \vec{s}_T\}$, the model with the largest likelihood function indicates the most likely microphone.

$$\sum_{t=1}^{T} \log p(\,\vec{s}_t|\lambda_k)\bar{M} \tag{8}$$

3. Data collection and test setup

Microphone identification experiments are conducted for different testing conditions: speaker-dependent, speaker-independent, and same content and different content training conditions. For this purpose, we use 3 different databases to determine the performance of the proposed microphone identification system. First, two 6-min-long speech utterances are recorded for each microphone, in a silent room, from a single speaker. One of the utterances has the same speech content for all of the microphones, reading from the same text, i.e. the first 4 pages of the famous novel Moby Dick, while the other utterance of each microphone has unique speech content, i.e. readings of different random pages from the same novel. These datasets are called DS1 and DS2. Hence, the speaker-dependent microphone identification performance is tested for 2 different training conditions, as same content and different content. Second, the TIMIT database is employed for speech samples recorded using different microphones. The TIMIT database consists of 6300 sentences, spoken by 192 females and 432 males. Each speaker speaks 10 phonetically rich sentences (2 dialect sentences, 5 phonetically compact sentences, and 3 phonetically diverse sentences), each of which is approximately 3-s long. The TIMIT database was recorded using a Sennheiser close-talking microphone at a rate of 16 kHz with a 16-bit sample resolution in wav format. For the speaker-independent microphone identification experiments, 40 speakers are selected from the test portion of the TIMIT database, and 120 phonetically diverse sentences of these 40 speakers are played and recorded by each microphone in the same room. This dataset is called DS3.

Sixteen different types of microphones (see Table 1) are tested, including headsets (behind-the-neck headset, over-the-head, and earbud headset), a lavalier microphone, and a desktop microphone. All of the microphones are based on the electret condenser-transducer–type and have 3.5-mm connectors to ensure that the recording is from the same acquisition device. Speech files (.wav) are created using microphones connected to a personal computer (Toshiba Satellite), at a sampling rate of 16 kHz and 16 bits per sample, with a mono channel in a quiet environment.

In all experiments, the same number feature vectors are used for each of the feature extraction methods. In other words, a 13-dimensional feature vector is extracted from the speech signal every 10 ms using a 25-ms Hamming window [25]. Moreover, the recordings are normalized prior to the feature extraction. For the GMM, a vector quantization algorithm is used for the initial parameter estimate and a diagonal covariance matrix is used to estimate the variances.

Table 1. Brands of the microphones used in the experiments.

Id	Brand	Model
M1	Sennheiser	PC-31
M2	A4Tech	HS-5P
M3	Creative	HS-350
M4	Enzatec	HS-903
M5	Genius	HS-500X
M6	Hama	CS-408
M7	Hama	NB-402
M8	Logitech	PC-120
M9	Microsoft	LX-2000
M10	Philips	SHM-1900
M11	Philips	SHM-1000
M12	Sony	DR-115DP
M13	Sony	ECM-DS70P
M14	Teac	HP-5S
M15	Trust	MC-1200
M16	Trust	Clip-17358

Each dataset is separated into 2 parts, as training and testing datasets. The GMM is trained for training speech durations of 30, 60, 90, 120, 150, and 180 s. Testing is carried out using different test utterance lengths (Tu) of 1 s and 3 s. Detailed information can be found in [13,15].

4. Experimental results

In this section, the experimental results obtained with the proposed method are presented (see Figure 1) to investigate the performance of our microphone recognition system. For each dataset, the MFCC, LPCC, and PLPC methods are used to analyze the impact of the feature extraction methods on the identification accuracy.

4.1. Results for DS1 dataset

Our first set of experiments aims to find out the microphone identification performance with the mixture components of the GMM under the same content training conditions. In order to do that, microphone models with 2-, 4-, 8-, 16-, 32-, 64-, and 128-component Gaussian densities are trained. The number of mixture components can be chosen experimentally. The recognition rates of the GMM-based system with different mixture component N values for the 3 different feature extraction methods are given in Table 2.

Table 2. The identification rates (%) of the GMM-based system for test utterance lengths of 1 s and 3 s (training duration of 180 s).

Mixtures (N)	MFCC		LPCC		PLPC	
	Tu = 1 s	Tu = 3 s	Tu = 1 s	Tu = 3 s	Tu = 1 s	Tu = 3 s
2	58.99	72.08	87.60	94.17	82.53	93.23
4	67.22	78.75	96.22	99.69	*89.27*	95.73
8	81.08	95.31	97.57	99.90	93.12	98.33
16	84.06	95.52	98.37	99.90	94.86	99.06
32	86.22	97.08	98.75	**100**	96.25	99.79
64	88.40	97.81	99.24	**100**	96.87	99.90
128	90.07	97.81	**99.44**	**100**	97.19	99.90

It can be seen from Table 2 that there is a sharp increase in the microphone identification performance from 2 to 16 Gaussian mixture components. The lower model order limit seems to be 16 mixture components to maintain good microphone identification performance. Above this lower model order limit, the microphone identification performance continues to increase at a lower rate with the number of mixture components for test utterance lengths of 1 s and 3 s. The results in Table 2 show that 16 mixtures is a good compromise between model complexity and identification performance. It is also shown that the longer the test utterance length is (1 s to 3 s) the more system performance increases. In addition, the LPCC method consistently outperforms the MFCC and PLPC methods over the different numbers of Gaussian mixtures.

In the next experiment, each microphone is modeled by a 16-component GMM trained using 30, 60, 90, 120, 150, and 180 s of speech. The microphone identification results using various amounts of training data are shown in Figure 3 for test utterance lengths of 1 s and 3 s.

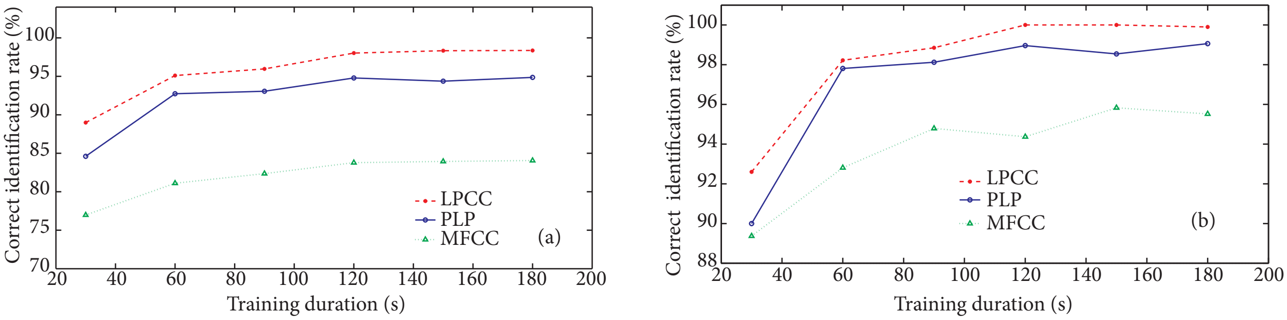

Figure 3. Microphone identification performance versus the training duration for the LPCC, PLPC, and MFCC methods, based on DS1: a) test utterance length of 1 s and b) test utterance length of 3 s.

It is shown in Figure 3 that the amount of training data has a strong impact on the microphone identification performance. As expected, the more training data used, the better the identification accuracy obtained. Furthermore, the LPCC method usually achieves better accuracies than the MFCC and PLPC methods for each of the training datasets used in this test.

4.2. Results for DS2 dataset

DS1 experiments are also carried out for different content using DS2. In this test, Table 3 tabulates the performance comparisons of the different feature extraction methods for test utterance lengths of 1 s and 3 s on DS2.

Table 3. The identification rates (%) of the GMM-based system for dataset DS2 (training duration of 180 s).

Mixtures (N)	MFCC		LPCC		PLPC	
	Tu = 1 s	Tu = 3 s	Tu = 1 s	Tu = 3 s	Tu = 1 s	Tu = 3 s
2	65.45	81.87	89.03	96.56	76.46	91.87
4	71.22	85.31	95.59	98.65	88.68	96.15
8	81.77	93.96	97.81	99.58	92.60	98.23
16	86.77	96.25	98.40	99.90	94.69	98.44
32	89.06	97.60	98.78	99.79	95.59	98.23
64	90.42	97.60	99.27	99.79	96.32	98.54
128	91.98	98.23	99.41	100	96.70	98.44

As can be seen from Table 3, the microphone identification rate increases with the increase in the mixture components. However, for mixture components of more than 16, the increase in the identification rate is slow. For the LPCC feature extraction method, the microphone identification rate of 1 s is about 99.41% and perfect accuracy can also be achieved for the testing utterance length of 3 s.

Moreover, the impact of the training durations is tested on DS2. For this purpose, a 16-component GMM and training durations of 30, 60, 90, 120, 150, and 180 s are used. Figure 4 summarizes the microphone identification performance with various amounts of training data and test utterance lengths for the LPCC, PLPC, and MFCC feature extraction techniques.

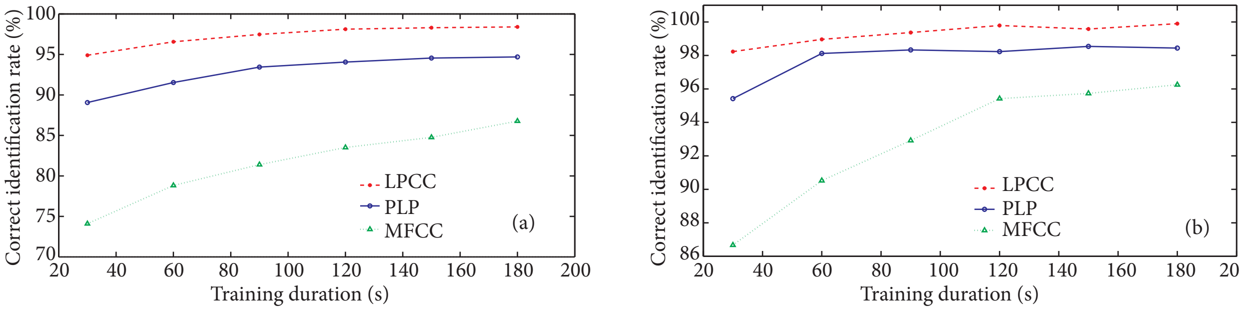

Figure 4. Microphone identification performance versus training duration for the LPCC, PLPC, and MFCC methods based on dataset DS2: a) test utterance length of 1 s and b) test utterance length of 3 s.

Figure 4 shows that the LPCC method has better performance than the other 2 methods under both conditions. The performance increases with more training data and a longer test utterance length.

The model order selection is also investigated for a smaller amount of training data. Table 4 shows the dependency of the model orders on the microphone identification performance for limited training data based on datasets DS1 and DS2. For each model order, the LPCC feature extraction method, a 1-s test utterance length, and 30 s of training data are used.

Table 4. Dependency of the model orders on the microphone identification performance (%) for a smaller amount of training data (30 s) based on datasets DS1 and DS2.

Mixtures (N)	2	4	8	16	32	64	128
DS1	76.56	81.53	86.28	88.99	90.52	90.35	90.00
DS2	82.67	90.45	93.26	94.90	95.48	95.90	95.31

As seen in Table 4, there are performance peaks at 32 components for dataset DS1 and 64 components for dataset DS2. As a result of these experiments, it is clearly seen that the amount of training data is an important factor for the success of the microphone identification system, since choosing the mixture components may be more difficult for smaller amounts of training data.

Additionally, we show the confusion matrix of the source microphone identification system on dataset DS2 for a 1-s test utterance length and 30 s of training data. The total number of test samples used in the experiment is 2880 (180 tests for each microphone). A confusion table of dataset DS2 for the LPCC feature extraction method is shown in Table 5.

4.3. Results for the DS3 dataset

Speaker-independent microphone identification is also studied. Microphone models are trained with speech data from 40 speakers. Microphone identification is performed for utterance lengths of 1 s and 3 s. The identification rates of the GMM-based system with different mixture component N values for the 3 different feature extraction methods are given in Table 6.

Table 5. Confusion matrix of the source microphone identification system on dataset DS2 for $N = 64$.

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16
M1	175	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0
M2	0	180	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M3	0	0	180	0	0	0	0	0	0	0	0	0	0	0	0	0
M4	0	0	0	179	0	0	0	0	0	0	0	0	0	0	1	0
M5	0	0	5	0	140	0	0	0	0	0	1	0	0	34	0	0
M6	0	0	0	0	0	179	0	0	0	1	0	0	0	0	0	0
M7	0	0	0	0	0	0	180	0	0	0	0	0	0	0	0	0
M8	0	0	0	0	0	0	0	180	0	0	0	0	0	0	0	0
M9	0	0	0	0	0	2	0	0	177	1	0	0	0	0	0	0
M10	0	0	0	0	0	0	0	0	0	180	0	0	0	0	0	0
M11	0	0	20	0	0	0	0	0	0	0	134	0	0	26	0	0
M12	0	0	0	0	0	0	0	0	0	0	0	180	0	0	0	0
M13	0	0	0	0	0	1	0	1	0	0	0	0	178	0	0	0
M14	0	0	0	0	9	0	0	0	0	0	8	0	0	163	0	0
M15	0	0	0	0	0	0	0	0	0	0	0	0	0	2	178	0
M16	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	179

Table 6. The identification rates (%) of GMM-based system for DS3 (training duration of 180 s).

Mixtures (N)	MFCC		LPCC		PLPC	
	Tu = 1 s	Tu = 3 s	Tu = 1 s	Tu = 3 s	Tu = 1 s	Tu = 3 s
2	80.00	92.19	88.30	96.04	71.28	84.17
4	79.79	88.65	95.66	99.27	81.35	90.52
8	87.05	94.37	96.04	99.48	87.78	94.79
16	88.89	95.73	96.46	99.06	89.69	96.04
32	89.97	96.56	97.60	99.58	91.22	96.67
64	91.84	96.87	97.92	99.58	92.05	97.18
128	91.28	96.87	98.19	99.58	92.33	97.5

In the last experiment, each microphone is modeled by a 16-component GMM trained using 30, 60, 90, 120, 150, and 180 s of speech. The microphone identification results using various amounts of training data are illustrated in Figure 5 for test utterance lengths of 1 s and 3 s, respectively.

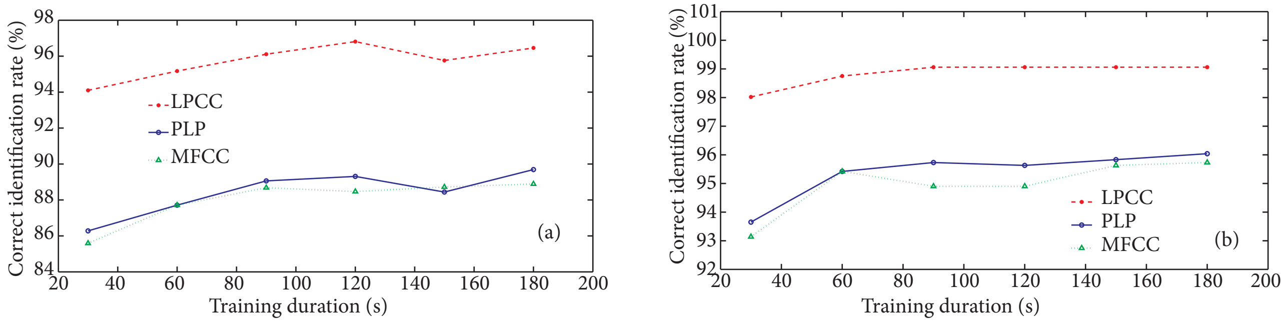

Figure 5. Microphone identification performance versus the training duration for the LPCC, PLPC, and MFCC methods based on dataset DS3: a) test utterance length of 1 s and b) test utterance length of 3 s.

Figure 5 shows that the MFCC and PLPC features performed approximately at the same level in the speaker-independent experiments for the 16-component GMM. It should be noted that the LPCC method achieves greater performance on discriminating microphones and it is an effective feature extraction technique for microphone identification.

5. Comparison with existing studies and discussions

The experimental results show that the LPCC features outperform the other features (MFCC and PLPC) for microphone identification. This is probably due to fact that the LPCC features are capable of capturing extra information from speech, i.e. increasing the ability to discriminate different microphones. Similar results can be seen in the field of language identification [27] and speaker recognition on the 2003 NIST SRE database [28].

It is generally seen that the results obtained with the speaker-dependent datasets (DS1 and DS2) are higher in the experiments with LPCC features. For example, the microphone identification results obtained with LPCC features ($N = 32$, $Tu = 3$ s, training duration is 180 s) is 100% for same content training condition, 99.79% for the different content training condition, and 99.58% for the speaker-independent condition (see Tables 2, 3, and 6). This is due to the fact that, in the speaker-dependent case, the vocal tract parameters are the same since one speaker is used.

Moreover, the confusion matrix, showing the differences in the microphone identification success rate between the individual microphones and consisting of the detailed identification results, is given in Table 3. The worst performance obtained for the recordings in Table 3 is observed in M5. This results from 34 tests from M5 mistakenly identified as M14. This is probably due to the fact that these 2 microphones have similar transducer properties, such as the size of the diaphragm, sensitivity, frequency response, and directionality.

In this section, the results obtained in this study are compared to the current results reported in the literature for microphone identification. In [10], time-domain features and MFCCs were employed in the microphone classification and the best classification accuracy of 75.99% was obtained with the naïve Bayes classifier. It was reported in [11] that, using linear regression models, a 93.5% accurate classification rate was obtained. In another study [12], using a decision tree and linear logistic regression models, a 100% classification rate was obtained for 4 and 7 microphones. The results in [13] were obtained using MFCCs and linear scale cepstral coefficients with the SVM classifier. They obtained classification accuracies of higher than 90% for landline telephone handsets and 8 different microphones. In a recent paper [14], the authors presented a context model for microphone forensics. They tested 74 supervised classification techniques and 8 clusters using data mining suite Waikato Environment for Knowledge Analysis (WEKA).

The test conditions of the current studies from the literature are different. For example, the training and test datasets were split by a rate of 66% for training and 34% for testing in [10], the other splitting strategy was a 10-fold cross-validation used in [11,14], reference samples were used in a rate of 80% to 20% for supervised training and testing in [12], and a 2-fold cross-validation setup was used in [13].

Since the datasets, the microphones, and the test conditions used in these works discussed above are different, it is not possible to make a comparison. Nevertheless, the identification rates of the proposed method and the existing results in the literature are presented in Table 7.

These promising results from this study must be recognized with their limitations. In the presented work, the studied speech recordings are conducted under restricted conditions, i.e. in a silent room and using 1 session. In real forensic cases, recordings may contain diverse types of noises or speaker movement; therefore, the effect on the speech quality of such cases in real forensic conditions should be kept in mind. Moreover, another important limitation of the study is that the recordings are collected from each microphone separately. It would produce better realistic results if all the recordings were collected simultaneously.

Table 7. The microphone identification rates (%) obtained by the proposed method and the other methods from the literature.

Author	Features-classifier	Number of microphones	Accuracy
Kraetzer et al. [10]	Statistical features and MFCC-naïve Bayes	4	75.99
Buchholz et al. [11]	Fourier coefficients - linear regression models	7	93.5
Kraetzer et al. [12]	Statistical features and MFCC-J48 and rank level fusion	4 7	100 100
Garcia-Romero and Espy-Wilson [13]	MFCC-GSV_SVM system for telephone handsets MFCC-GSV_SVM system for microphone s	8 8	93.2 99.0
Kraetzer et al. [14]	AMSL audio feature extractor -Weka classifiers for parallel recording	4	82.5
This work	LPCC-GMM (speaker-dependent) LPCC-GMM (speaker-independent)	16 16	100 99.58

6. Conclusion

In this paper, we investigate the performance of a source microphone identification system using speech samples. Three different feature extraction methods (i.e. LPCC, PLPC, and MFCC) and a GMM-based modeling technique are employed to identify the source microphones. The system is tested for speaker-dependent (DS1 and DS2) and speaker-independent (DS3) conditions, using 16 different microphones. The results show that for all of the recording sets, the LPCC feature extraction technique achieves the highest identification rate.

In addition, it is aimed to find out the important parameters that affect the microphone identification performance. The experimental results show that the microphone identification performance largely depends on the number of mixture components of the GMM, amount of training data, and the test utterance length.

In future, open-set experiments will be performed to test the system under a wide variety of noise conditions with more realistic speech data, and the system will be improved to use different feature selection methods and classifiers.

References

[1] S. Bayram, H.T. Sencar, N. Memon, İ. Avcıbaş, "Source camera identification based on CFA interpolation", Proceedings of the IEEE International Conference on Image Processing, Vol. 3, pp. 69–72, 2005.

[2] J. Lukas, J. Fridrich, M. Goljan, "Digital camera identification from sensor pattern noise", IEEE Transactions on Information Forensics and Security, Vol. 1, pp. 205–214, 2006.

[3] H. Farid, "Detecting digital forgeries using bispectral analysis", Technical Report Perceptual Science Group, Massachusetts Institute of Technology, 1999.

[4] M. Kajstura, A. Trawinska, J. Hebenstreit, "Application of the electrical network frequency (ENF) criterion. A case of a digital recording", Forensic Science International, Vol. 155, pp. 165–171, 2005.

[5] C. Grigoras, "Applications of ENF criterion in forensic audio, video, computer, and telecommunication analysis", Forensic Science International, Vol. 167, pp. 136–145, 2007.

[6] D.P. Nicolalde, J.A. Apolinário, L.W.P. Biscainho, "Audio authenticity: detecting ENF discontinuity with high precision phase analysis", IEEE Transactions on Information Forensics and Security, Vol. 5, pp. 534–543, 2010.

[7] R. Yang, Q. Zhenhua, H. Jiwu, "Detecting digital audio forgeries by checking frame offsets", Proceedings of the 10th ACM workshop on Multimedia and Security, pp. 21–26, 2008.

[8] L. Burget, P. Matejka, P. Schwarz, O. Glembek, J. Cernocký, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, pp. 1979–1986, 2007.

[9] D.A. Reynolds, "HTIMIT and LLHDB: speech corpora for the study of handset transducer effects", IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 1535–1538, 1997.

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

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

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

[13] D. Garcia Romero, C.Y. Espy Wilson, "Automatic acquisition device identification from speech recordings", Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, pp. 1806–1809, 2010.

[14] C. Kraetzer, K. Qian, M. Schott, J. Dittmann, "A context model for microphone forensics and its application in evaluations", Media Watermarking, Security, and Forensics III, Vol. 7880, 2011.

[15] C. Hanilçi, F. Ertaş, T. Ertaş, Ö. Eskidere, "Recognition of brand and models of cell-phones from recorded speech signals", IEEE Transactions on Information Forensics and Security, Vol. 7, pp. 625–634, 2012.

[16] J.L. Flamagan, "A singular advance in conversion of acoustic signals to electrical form: the electret microphone", IEEE Signal Processing Magazine, Vol. 27, pp. 102–116, 2010.

[17] W.J. Wang, R.M. Lin, Q.B. Zhou, X.X. Li, "Modeling and characterization of a silicon condenser microphone", Journal of Micromechanics and Microengineering, Vol. 14, pp. 403–409, 2004.

[18] P.R. Scheeper, A.G.H. Van der Donk, W. Olthuis, P. Bergveld, "A review of silicon microphones", Sensors and Actuators A: Physical, Vol. 44, pp. 1–11, 1994.

[19] P.C. Hsu, C.H. Mastrangelo, K.D. Wise, "A high sensitivity polysilicon diaphragm condenser microphone", Proceedings of the IEEE Workshop on Micro Electro Mechanical Systems, pp. 580–585, 1998.

[20] Q. Zou, Z. Tan, Z. Wang, J. Pang, X. Qian, Q. Zhang, R. Lin, S. Yi, H. Gong, L. Liu, Z. Li, "A novel integrated silicon capacitive microphone-floating electrode 'electret' microphone (FEEM)", Journal of Microelectromechanical Systems, Vol. 7, pp. 224–234, 1998.

[21] L.R. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, New Jersey, Prentice Hall, 1993.

[22] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", Journal of the Acoustical Society of America, Vol. 87, 1738–1752, 1990.

[23] P. Melmerstein, S. Davis, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 28, pp. 357–336, 1980.

[24] R. Mammone, X. Zhang, R. Ramachandran, "Robust speaker recognition: a feature-based approach", IEEE Signal Processing Magazine, Vol. 13, pp. 58–71, 1996.

[25] M. Slaney, Auditory Toolbox: A MATLAB Toolbox for Auditory Modeling, Work Technical Report, Interval Research Corporation, pp. 29–32, 1998.

[26] D.A. Reynolds, R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Transactions on Speech Audio Processing, Vol. 3, pp. 72–83, 1995.

[27] E. Wong, S. Sridharan, "Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification", International Symposium on Intelligent Multimedia, Video and Speech Processing, pp. 95–98, 2002.

[28] W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, P.A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition", Computer Speech and Language, Vol. 20, pp. 210–229, 2006.