

# Mobile Phone Identification from Recorded Speech Signals Using Non-Speech Segments and Universal Background Model Adaptation

Dimitrios Kritsiolis<sup>a</sup> and Constantine Kotropoulos<sup>b</sup>

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece

**Keywords:** Mobile Phone Identification, Digital Speech Processing, Audio Forensics, Voice Activity Detection, Universal Background Model.

**Abstract:** Mobile phone identification from recorded speech signals is an audio forensic task that aims to establish the authenticity of a speech recording. The typical methodology to address this problem is to extract features from the entire signal, model the distribution of the features of each phone, and then perform classification on the testing data. Here, we demonstrate that extracting features from non-speech segments or extracting features from the entire recording and modeling them using a Universal Background Model (UBM) of speech improves classification accuracy. The paper's contribution is in the disclosure of experimental results on two benchmark datasets, the MOBIPHONE and the CCNU Mobile datasets, demonstrating that non-speech features and UBM modeling yield higher classification accuracy even under noisy recording conditions and amplified speaker variability.

## 1 INTRODUCTION


Mobile phone identification from recorded speech signals aims to establish the authenticity of a speech recording. It is not uncommon for speech recordings from mobile phones to be presented as evidence in criminal investigations or other legal proceedings. However, for the recording to be admitted as evidence in a legal case, the court must be convinced of its authenticity and integrity (Maher, 2009).


Here, we examine the problem of brand and model identification of mobile phones from recorded speech signals. The term brand refers to the mobile phone manufacturer, e.g. LG, and the term model refers to a specific product from a manufacturer, e.g. LG L3.

A three-stage process can describe mobile phone identification from recorded speech signals. In the first stage, features capable of capturing the intrinsic trace each device leaves on the speech signal are extracted. In the second stage, the extracted features are used to train models that represent the distribution of features of each mobile phone. In the third stage, classification is performed using the trained models and the extracted features. The features to be examined are the Mel Frequency Cepstral Coef-

ficients (MFCCs) (Davis and Mermelstein, 1980) extracted on a frame basis. The extracted feature vectors are then used to train Gaussian Mixture Models (GMMs) using the Expectation Maximization (EM) algorithm (Dempster et al., 1977). The GMMs can be used for Maximum Likelihood (ML) classification or the extraction of fixed dimensional feature vectors, known as Gaussian Supervectors (GSVs) (Garcia-Romero and Espy-Wilson, 2010). The GSVs are obtained by concatenating the mean vectors (and optionally the main diagonals of the covariance matrices) of the GMMs. They are used in Support Vector Machine (SVM) classification.

We will explore two different approaches to mobile phone identification. The first approach is to extract the features from the non-speech segments of the signal since the speech parts might introduce speaker-related noise into the device's intrinsic trace. The second approach stems from the fact that mobile phone identification based on recorded speech signals is similar to the speaker recognition problem. Thus, approaches used in the latter field can be applied in the former. One of these approaches is the Universal Background Model (UBM) of speech (Reynolds et al., 2000) that is employed to adapt the mobile phones' feature vectors by means of Maximum A Posteriori (MAP) adaptation. GSVs are then extracted from the UBM-adapted GMMs and are used

<sup>a</sup>  <https://orcid.org/0009-0009-3845-5928>

<sup>b</sup>  <https://orcid.org/0000-0001-9939-7930>

in the classification stage.

The paper contributes to disclosing experimental evidence using either ML, SVM, or neural classification on two benchmark datasets, namely the MOBIPHONE dataset and the CCNU Mobile dataset. The empirical results demonstrate that both the non-speech and the UBM approaches yield higher accuracy than the traditional approach of extracting features from the entire speech signal and modeling them with GMMs trained via the EM algorithm, even under noisy recording conditions and amplified speaker variability.

## 2 RELATED WORK

(Hanilci et al., 2011) proposed a mobile phone identification system using MFCCs, Vector Quantization classifiers, and SVMs. (Hanilçi and Kinnunen, 2014) extended their previous work by showing that MFCCs obtained from the non-speech segments yield higher mutual information than the MFCCs obtained from the speech segments or the entire speech recording.

(Kotropoulos and Samaras, 2014) used MFCCs to train GMMs and to obtain GSVs, which were used with SVMs, a Radial Basis Function neural network, and a Multilayer Perceptron. To test their methods, they collected the MOBIPHONE dataset. (Kotropoulos, 2014) obtained sketches of features from large-size raw feature vectors and used them with a Sparse Representation Classifier and SVMs on the Lincoln-Labs Handset Database (Reynolds, 1997) and on the MOBIPHONE dataset.

(Baldini and Amerini, 2019) proposed a mobile phone identification approach by stimulating the phones' microphones with non-voice sounds. They extracted the frequency spectrum magnitude and used it as the input to a Convolutional Neural Network (CNN), which performed the identification task. (Baldini and Amerini, 2022) continued their work by obtaining spectral entropy features based on Shannon entropy and Renyi entropy to implement dimensionality reduction of the spectral representation of audio signals. These features were then the input to their previous CNN.

(Giganti et al., 2022) proposed identifying mobile phones from their microphone under noisy conditions by applying convolutional neural network-based denoising on the signal's spectrogram.

(Berdich et al., 2022) explored mobile phone microphone fingerprinting based on human speech, environmental sounds, and several live recordings performed outdoors, using supervised machine learning methods.

(Zeng et al., 2020) proposed an end-to-end deep source recording device identification system for web media forensics utilizing the MFCCs, the GSVs, and the i-vectors. (Zeng et al., 2023) expanded their work by proposing a spatio-temporal representation learning framework utilizing the MFCCs and the GSVs.

(Berdich et al., 2023) surveyed smartphone fingerprinting technologies based on sensor characteristics.

(Qamhan et al., 2023) proposed a transformer network for authenticating the source microphone. The Audio Forensic Dataset for Digital Multimedia Forensics (Khan et al., 2018) and the King Saud University speech database (Alsulaiman et al., 2013) were used.

(Leonzio et al., 2023) addressed audio splicing detection and localization by using the CNN in (Baldini and Amerini, 2019) for mobile phone classification. They applied  $K$ -means to extracted features from intermediate layers of the CNN to detect splicing and employed the cosine distance for localization between consecutive frame vectors.

## 3 MOBILE PHONE IDENTIFICATION

Due to tolerances in the manufacturing of electronic components, each implementation of an electronic circuit, particularly when a microphone is included, cannot have the same transfer function (Hanilci et al., 2011). The recorded speech signal's frequency spectrum can be considered as the multiplication of the spectrum of the original speech signal with the transfer function of the mobile phone's microphone. Consequently, every mobile phone leaves its own device fingerprint on the recorded speech signal.

Mobile phone identification is very similar to closed-set text-independent speaker recognition. A simple block diagram of an automatic speaker recognition system is shown in Fig. 1. The same principles used in speaker recognition also apply to mobile phone identification. Closed-set mobile phone identification is done in 3 stages: feature extraction, modeling, and classification.

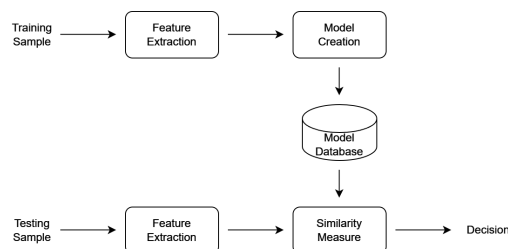


Figure 1: Block diagram of a basic closed-set automatic speaker recognition system.

The MFCCs are the features used and they are based on the speech signal spectrum. The shape of the vocal tract gives us information about the speaker uttering the speech or the phoneme being uttered. It is found in the spectral envelope of the short-time power spectrum. The MFCCs try to represent that envelope.

Once the MFCCs have been extracted, each mobile phone model or brand is modeled using a GMM trained on the corresponding short-time features using the EM algorithm. The GMM probability density function is used to measure the similarity between the testing data of an unknown mobile phone and the trained GMMs. To obtain a fixed-dimensional vector representing a speech utterance for SVM and neural classification, we construct GSVs by concatenating the parameters of a GMM trained on the speech utterance.

After training the GMMs and extracting the GSVs, classification is performed. Firstly, an ML classifier is utilized by using the MFCC vectors extracted from the testing samples and comparing their similarity with the trained GMMs. SVMs are also trained using the GSVs extracted from the GMMs. Finally, we employ the neural architecture in (Zeng et al., 2023) for mobile phone identification that considers both the MFCCs and the GSVs. The mobile phone identification process outlined above will be referred to as the traditional approach. We examine two different modifications to the traditional approach.

The first modification is referred to as the non-speech approach. Speech parts have speaker-dependent information that is irrelevant to the task at hand. The noise-like signals in the non-speech parts have a flatter spectral density and better capture the mobile phone's recording circuitry transfer function. Accordingly, the feature extraction stage is modified to extract MFCCs from non-speech parts. To obtain non-speech parts, we use either a voice activity detection algorithm (Sohn et al., 1999) or speech enhancement to estimate a clean speech signal and subtract it from the initial speech signal to obtain a residual noise signal. The speech enhancement method utilizes the noise estimation method described in (Martin, 2001).

The second modification, referred to as the UBM approach, stems from the fact that the problem of mobile phone identification from recorded speech signals is very similar to the speaker recognition problem (Hanilci et al., 2011). A speech signal is modeled in the time domain as the convolution of the vocal cords' excitation and the vocal tract's impulse response. Speaker recognition aims to extract features that can distinguish the transfer function of the vocal tract of each speaker. A recorded speech signal, however, also contains a convolutional distortion in-

duced by the mobile phone's microphone. Suppose that the recorded speech signal is produced by a transfer function given by the serial composition of the vocal tract transfer function and the microphone's transfer function. In that case, the task at hand becomes tantamount to speaker recognition. Thus, the use of a UBM model, a large speaker-independent GMM trained via the EM algorithm to capture the distribution of the MFCCs in a very large set of speakers, becomes justified. The UBM is used to adapt the GMMs directly via MAP adaptation.

## 4 DATASETS

### 4.1 MOBIPHONE and Augmentations

The MOBIPHONE dataset (Kotropoulos and Samaras, 2014) consists of recorded speech utterances of 24 speakers, 12 males and 12 females, using 21 mobile phones belonging to 7 brands. The 24 speakers were randomly chosen from the TIMIT dataset (Garofolo et al., 1988). Each speaker uttered 10 sentences approximately 3 seconds long. Each speaker's sentences were concatenated in a single WAV file. Thus, the dataset contains 504 WAV files sampled at 16kHz, one for each speaker recorded on each mobile phone.

To compare the performance of the different feature extraction and modeling approaches, 8 augmentations were applied to the entire dataset to create 8 additional versions of the MOBIPHONE dataset.

The first set of 3 augmentations aims to simulate different recording conditions. It includes the addition of Gaussian noise at a random signal-to-noise ratio (SNR) between 10-20 dB, the addition of background chattering noise at a random SNR between 10-20 dB, and the addition of reverberation.

The second set of another 5 augmentations aims to introduce more speaker variability by the removal of random croppings from the audio signal to enlarge the dataset, the adjustment of the loudness volume of some segments (i.e., multiplication by a constant in the range  $[0.5, 2]$ ), the modification of the pitch with a factor in the range  $[-10, 10]$ , the modification of the speed (i.e., time scaling with a factor in the range  $[0.5, 2]$ ), and Vocal Tract Length Perturbation (VTLP) with a factor in the range  $[0.9, 1.1]$ .

### 4.2 CCNU Mobile

The Central China Normal University (CCNU) Mobile dataset (Zeng et al., 2020) consists of speech recordings from the TIMIT dataset recorded by 45

mobile phones belonging to 9 brands. Multiple devices of the same mobile phone model are included in the dataset. The recordings of each phone consist of 642 audio samples per mobile phone with a duration of about 8 seconds sampled at 32kHz. Unfortunately, the entire dataset is not publicly available. We were able to use a small subset of the dataset, which consists of 5 8-second-long audio samples of female speakers per mobile phone. This data scarcity will allow us to test our methods under realistic audio forensic conditions with limited data.

## 5 EXPERIMENTAL RESULTS

### 5.1 Results on MOBIPHONE

Experiments were conducted on the MOBIPHONE dataset using the traditional, non-speech, and UBM approaches. Two tasks were considered, namely brand identification and model identification. We used a 50%/50% train/test split with half male and half female speakers in both sets.

23 MFCCs were extracted using 20 ms Hamming windows with 10 ms overlap. The MFCCs for the non-speech approach were extracted from non-speech frames obtained via voice activity detection. The MFCCs for the traditional and UBM approaches are referred to as audio MFCCs, while the MFCCs for the non-speech approach are referred to as non-speech MFCCs.

Model- and brand-dependent GMMs were trained via the EM algorithm, using a `k-means++` initialization for both the audio and non-speech MFCC vectors. We trained GMMs with the number of components in the range 1, 2, 4, 8, 16, 32, and 64. These GMMs were used in the ML classification. For the SVM classification, utterance-dependent GMMs were trained using audio and non-speech MFCC vectors. The components of these GMMs were in the range 1, 2, and 4 because the classification accuracy was drastically reduced when more than 4 components were employed in the extracted GSVs due to the large number of features.

Finally, utterance-dependent GMMs were also trained using UBM MAP adaptation applied to the audio MFCCs. The UBMs were trained on the TIMIT dataset. In the case of the UBM-adapted GSVs, it was seen that SVMs with UBMs up to 16 components yielded good results. As is most common, only the means were adapted.

The results of the initial experiments on the MOBIPHONE dataset are listed in Tables 1-4. The first column shows the number of components of the

GMMs. The remaining columns display the accuracies of the traditional approach (Audio), the non-speech approach (Non-speech), and the UBM approach (UBM) on the corresponding classification method.

Table 1: ML classification accuracy for brand identification on the MOBIPHONE dataset using audio and non-speech MFCCs.

| # Components | Audio    | Non-speech |
|--------------|----------|------------|
| 1            | 66.6667% | 79.3651%   |
| 2            | 73.4127% | 86.1111%   |
| 4            | 88.4921% | 96.8254%   |
| 8            | 91.6667% | 99.2063%   |
| 16           | 88.0952% | 99.6032%   |
| 32           | 93.254%  | 100%       |
| 64           | 98.0159% | 99.6032%   |

Table 2: ML classification accuracy for model identification on the MOBIPHONE dataset using audio and non-speech MFCCs.

| # Components | Audio    | Non-speech |
|--------------|----------|------------|
| 1            | 96.0317% | 98.8095%   |
| 2            | 98.8095% | 99.6032%   |
| 4            | 98.4127% | 100%       |
| 8            | 99.2063% | 100%       |
| 16           | 98.4127% | 100%       |
| 32           | 98.8095% | 100%       |
| 64           | 98.4127% | 100%       |

A first remark on the results in Tables 1-4 is that the brand identification problem seems harder than the model identification problem. In practice, the brand can be easily deduced from the model. Since model identification performs better, brand identification is of no practical use. However, by examining it we see the accuracy increase offered by non-speech MFCCs and UBM adaptation. The non-speech and the UBM approaches perform better than the traditional approach when performing ML or SVM classification. In the non-speech approach, we also note that the GSVs with the covariance don't always yield an accuracy increase compared to the mean GSVs. This can also be verified in the following experiments.

The same classification procedure was followed on the augmented versions of the MOBIPHONE dataset. The aim is to demonstrate how the different approaches perform under different types of noise. The 504 samples of the baseline MOBIPHONE dataset were expanded by the 504 samples of one or more augmented versions.

The best accuracy on the first set of augmentations, which simulate different recording conditions,

Table 3: SVM classification accuracy for brand identification on the MOBIPHONE dataset using GSVs extracted from audio MFCCs, non-speech MFCCs, and UBM adaptation.

| # Components | Audio    |           | Non-speech |           | UBM      |
|--------------|----------|-----------|------------|-----------|----------|
|              | Means    | Means+Cov | Means      | Means+Cov | Means    |
| 1            | 93.6508% | 96.4286%  | 97.619%    | 97.619%   | 98.0159% |
| 2            | 93.6508% | 96.0317%  | 97.619%    | 97.619%   | 97.2222% |
| 4            | 82.5397% | 85.3175%  | 92.8571%   | 92.8571%  | 98.8095% |
| 8            | -        | -         | -          | -         | 98.8095% |
| 16           | -        | -         | -          | -         | 96.0317% |

Table 4: SVM classification accuracy for model identification on the MOBIPHONE dataset using GSVs extracted from audio MFCCs, non-speech MFCCs, and UBM adaptation.

| # Components | Audio    |           | Non-speech |           | UBM      |
|--------------|----------|-----------|------------|-----------|----------|
|              | Means    | Means+Cov | Means      | Means+Cov | Means    |
| 1            | 94.0476% | 96.4286%  | 99.2063%   | 98.8095%  | 94.0476% |
| 2            | 94.0476% | 96.4286%  | 97.619%    | 97.2222%  | 98.0159% |
| 4            | 92.8571% | 94.4444%  | 96.4286%   | 95.2381%  | 98.8095% |
| 8            | -        | -         | -          | -         | 98.8095% |
| 16           | -        | -         | -          | -         | 94.4444% |

Table 5: The best classification accuracy on MOBIPHONE for brand identification on the first set of augmentations.

|               | Audio                  |                          | Non-speech             |                          | UBM                     |
|---------------|------------------------|--------------------------|------------------------|--------------------------|-------------------------|
|               | ML                     | SVM                      | ML                     | SVM                      | SVM                     |
| Baseline      | 98.0159% <sup>64</sup> | 96.4286% <sup>1/MC</sup> | 100% <sup>32</sup>     | 97.619% <sup>1/M</sup>   | 98.8095% <sup>4/M</sup> |
| Gaussian      | 85.7143% <sup>64</sup> | 89.2857% <sup>1/M</sup>  | 81.3492% <sup>64</sup> | 92.2619% <sup>1/MC</sup> | 89.2857% <sup>1/M</sup> |
| Background    | 92.0635% <sup>64</sup> | 92.8571% <sup>2/M</sup>  | 86.5079% <sup>64</sup> | 91.6667% <sup>1/M</sup>  | 98.4127% <sup>4/M</sup> |
| Reverberation | 96.8254% <sup>32</sup> | 95.0397% <sup>1/MC</sup> | 99.0079% <sup>64</sup> | 97.2222% <sup>1/MC</sup> | 98.4127% <sup>8/M</sup> |
| All           | 88.8889% <sup>64</sup> | 93.0556% <sup>1/MC</sup> | 83.3333% <sup>64</sup> | 90.4762% <sup>1/M</sup>  | 95.2381% <sup>2/M</sup> |

Table 6: The best classification accuracy on MOBIPHONE for model identification on the first set of augmentations.

|               | Audio                  |                          | Non-speech             |                          | UBM                     |
|---------------|------------------------|--------------------------|------------------------|--------------------------|-------------------------|
|               | ML                     | SVM                      | ML                     | SVM                      | SVM                     |
| Baseline      | 99.2063% <sup>8</sup>  | 96.4286% <sup>1/MC</sup> | 100% <sup>4</sup>      | 99.2063% <sup>1/M</sup>  | 98.8095% <sup>4/M</sup> |
| Gaussian      | 89.881% <sup>16</sup>  | 90.6746% <sup>1/MC</sup> | 78.9683% <sup>64</sup> | 93.8492% <sup>1/M</sup>  | 92.8571% <sup>2/M</sup> |
| Background    | 93.0556% <sup>16</sup> | 95.4365% <sup>1/MC</sup> | 83.3333% <sup>64</sup> | 94.0476% <sup>1/MC</sup> | 97.4206% <sup>2/M</sup> |
| Reverberation | 98.6111% <sup>8</sup>  | 96.2302% <sup>2/MC</sup> | 99.6032% <sup>32</sup> | 98.4127% <sup>1/M</sup>  | 98.2143% <sup>4/M</sup> |
| All           | 91.5675% <sup>64</sup> | 92.2619% <sup>1/M</sup>  | 80.2579% <sup>64</sup> | 91.8651% <sup>1/MC</sup> | 96.2302% <sup>2/M</sup> |

is summarized in Tables 5-6 for brand and model identification, respectively. The best hyperparameters are indicated, namely the number of Gaussian components and the type of GSVs (means/means and covariance). The best accuracy listed in Tables 1-4 is quoted in the row Baseline of Tables 5-6.

For brand and model identification, non-speech ML classification performs worse than the traditional ML classification in all cases except for the reverberation augmentation. However, non-speech SVM classification performs better than the traditional SVM classification in the Gaussian noise and reverberation augmentations. In all cases, the UBM-adapted SVM classification achieves higher accuracy than the traditional SVM classification. It also yields the top accuracy when all the augmentations are used together.

Overall, the UBM approach performs better on the first set of augmentations. This is attributed to the probabilistic alignment of the MFCC vectors to the UBM, which can mitigate the effects of the noise on the features.

The best classification accuracy on the second set of augmentations is quoted in Tables 7-8 for brand and model identification, respectively. The non-speech ML classification of brand and model achieves the highest accuracy for all augmentations but the VTLP augmentation in the model identification task. Speaker variability is eliminated because the non-speech features are not affected much by the speaker-dependent noise amplified with this set of augmentations. When SVM classification is employed, the UBM and non-speech approaches achieve better ac-

Table 7: The best classification accuracy on MOBIPHONE for brand identification on the second set of augmentations.

|          | Audio                  |                          | Non-speech             |                          | UBM                      |
|----------|------------------------|--------------------------|------------------------|--------------------------|--------------------------|
|          | ML                     | SVM                      | ML                     | SVM                      | SVM                      |
| Baseline | 98.0159% <sup>64</sup> | 96.4286% <sup>1/MC</sup> | 100% <sup>32</sup>     | 97.619% <sup>1/M</sup>   | 98.8095% <sup>4/M</sup>  |
| Crop     | 98.2143% <sup>64</sup> | 96.8254% <sup>1/MC</sup> | 99.8016% <sup>64</sup> | 98.6111% <sup>1/M</sup>  | 99.0079% <sup>16/M</sup> |
| Loudness | 97.4206% <sup>64</sup> | 96.0317% <sup>1/MC</sup> | 99.0079% <sup>8</sup>  | 99.2063% <sup>1/M</sup>  | 97.8175% <sup>2/M</sup>  |
| Pitch    | 92.0635% <sup>64</sup> | 86.3095% <sup>1/MC</sup> | 97.4206% <sup>64</sup> | 94.8413% <sup>2/MC</sup> | 94.4444% <sup>16/M</sup> |
| Speed    | 98.2143% <sup>64</sup> | 97.0238% <sup>2/MC</sup> | 99.6032% <sup>32</sup> | 99.2063% <sup>1/MC</sup> | 98.4127% <sup>16/M</sup> |
| VTLP     | 96.4286% <sup>64</sup> | 98.0159% <sup>1/MC</sup> | 98.4127% <sup>32</sup> | 97.619% <sup>1/MC</sup>  | 97.619% <sup>8/M</sup>   |
| All      | 95.6349% <sup>64</sup> | 95.172% <sup>1/MC</sup>  | 98.0159% <sup>64</sup> | 96.627% <sup>1/MC</sup>  | 96.9577% <sup>8/M</sup>  |

Table 8: The best classification accuracy on MOBIPHONE for model identification on the second set of augmentations.

|          | Audio                  |                          | Non-speech             |                          | UBM                     |
|----------|------------------------|--------------------------|------------------------|--------------------------|-------------------------|
|          | ML                     | SVM                      | ML                     | SVM                      | SVM                     |
| Baseline | 99.2063% <sup>8</sup>  | 96.4286% <sup>1/MC</sup> | 100% <sup>4</sup>      | 99.2063% <sup>1/M</sup>  | 98.8095% <sup>4/M</sup> |
| Crop     | 99.4048% <sup>32</sup> | 96.2302% <sup>2/MC</sup> | 100% <sup>4</sup>      | 99.0079% <sup>1/MC</sup> | 99.6032% <sup>4/M</sup> |
| Loudness | 98.8095% <sup>8</sup>  | 97.4206% <sup>1/MC</sup> | 99.2063% <sup>8</sup>  | 99.0079% <sup>1/M</sup>  | 99.0079% <sup>2/M</sup> |
| Pitch    | 94.6429% <sup>64</sup> | 86.9048% <sup>1/MC</sup> | 98.4127% <sup>32</sup> | 97.0238% <sup>1/MC</sup> | 97.4206% <sup>4/M</sup> |
| Speed    | 98.8095% <sup>8</sup>  | 96.627% <sup>1/MC</sup>  | 100% <sup>4</sup>      | 99.4048% <sup>1/MC</sup> | 98.0159% <sup>8/M</sup> |
| VTLP     | 99.4048% <sup>64</sup> | 98.0159% <sup>1/MC</sup> | 98.8095% <sup>4</sup>  | 99.2063% <sup>1/MC</sup> | 98.8095% <sup>4/M</sup> |
| All      | 97.2884% <sup>64</sup> | 93.1217% <sup>1/M</sup>  | 98.8095% <sup>32</sup> | 96.6931% <sup>1/M</sup>  | 98.8095% <sup>8/M</sup> |

Table 9: The best classification accuracy on MOBIPHONE for brand identification on both sets of augmentations combined.

|                   | Audio                  |                          | Non-speech            |                         | UBM                     |
|-------------------|------------------------|--------------------------|-----------------------|-------------------------|-------------------------|
|                   | ML                     | SVM                      | ML                    | SVM                     | SVM                     |
| Baseline          | 98.0159% <sup>64</sup> | 96.4286% <sup>1/MC</sup> | 100% <sup>32</sup>    | 97.619% <sup>1/M</sup>  | 98.8095% <sup>4/M</sup> |
| All augmentations | 91.4021% <sup>64</sup> | 91.9312% <sup>1/MC</sup> | 89.903% <sup>64</sup> | 93.9594% <sup>1/M</sup> | 93.6067% <sup>2/M</sup> |

Table 10: The best classification accuracy on MOBIPHONE for model identification on both sets of augmentations combined.

|                   | Audio                 |                          | Non-speech             |                         | UBM                     |
|-------------------|-----------------------|--------------------------|------------------------|-------------------------|-------------------------|
|                   | ML                    | SVM                      | ML                     | SVM                     | SVM                     |
| Baseline          | 99.2063% <sup>8</sup> | 96.4286% <sup>1/MC</sup> | 100% <sup>4</sup>      | 99.2063% <sup>1/M</sup> | 98.8095% <sup>4/M</sup> |
| All augmentations | 93.739% <sup>64</sup> | 92.0635% <sup>2/MC</sup> | 89.9471% <sup>64</sup> | 94.0035% <sup>1/M</sup> | 93.2099% <sup>2/M</sup> |

Table 11: Mean and standard deviation of the classification accuracy for brand identification on the neural network using Zeng’s UBM.

|               | Zeng’s UBM         |
|---------------|--------------------|
| MFCC branch   | 41.7777% (5.1225%) |
| GSV branch    | 91.4074% (0.9999%) |
| Fusion branch | 70.7407% (7.7%)    |

Table 12: Mean and standard deviation of the classification accuracy for model identification on the neural network using Zeng’s UBM.

|               | Zeng’s UBM          |
|---------------|---------------------|
| MFCC branch   | 23.9555% (6.6566%)  |
| GSV branch    | 96.2222% (1.54%)    |
| Fusion branch | 73.8888% (10.2003%) |

curacy than the traditional approach.

When both sets of augmentations are used together, the brand and model identification accuracy is summarized in Tables 9-10. In both tasks, non-speech SVMs yield the best accuracy. Non-speech ML classification performs worse than the traditional

ML classification. UBM-adapted SVM classification also performs better than the traditional SVM classification.

## 5.2 Results on CCNU Mobile

The CCNU Mobile dataset was used to compare the identification methods further. In addition to the ML and SVM classification methods, the neural architecture designed specifically for mobile phone identification in (Zeng et al., 2023) was included.

The entirety of the CCNU Mobile dataset is not available, and we could use only a small sample of it containing 5 audio samples from female speakers for each of the 45 mobile phones. However, the limited data helps us test the quality of the extracted data since the feature vectors are so few the classifiers have to rely on the data’s quality and not quantity to yield better results.

The first two speakers were chosen for the train set, while the remaining 3 speakers were selected for the test set. To be consistent with the features used

Table 13: Mean and standard deviation of the classification accuracy for brand identification using the three approaches with ML, SVMs, and the neural network proposed by (Zeng et al., 2023).

|               | Audio               | Non-speech         | UBM                |
|---------------|---------------------|--------------------|--------------------|
| ML-64         | 86.5899% (1.7982%)  | 99.2593% (0%)      | -                  |
| SVM-1M        | 86.0741% (4.2881%)  | 86.8148% (5.8514%) | -                  |
| SVM-1MC       | 68.6667% (10.0482%) | 83.4074% (4.7263%) | -                  |
| SVM-4M        | -                   | -                  | 86.5926% (1.3726%) |
| MFCC branch   | 45.3333% (6.5309%)  | 60.3703% (6.6369%) | 44.5185% (3.7357%) |
| GSV branch    | 61.0370% (3.7803%)  | 83.3333% (2.5241%) | 94.7407% (1.3725%) |
| Fusion branch | 60.5926% (4.9481%)  | 81.6296% (3.9009%) | 70.0740% (7.3761%) |

Table 14: Mean and standard deviation of the classification accuracy for model identification using the three approaches with ML, SVMs, and the neural network proposed by (Zeng et al., 2023).

|               | Audio              | Non-speech         | UBM                 |
|---------------|--------------------|--------------------|---------------------|
| ML-64         | 98.5185% (0%)      | 98.5185% (0%)      | -                   |
| SVM-1M        | 90.8148% (5.7292%) | 96.5185% (3.5447%) | -                   |
| SVM-1MC       | 81.9259% (6.3356%) | 90.7407% (1.0030%) | -                   |
| SVM-4M        | -                  | -                  | 93.3333% (0.7808%)  |
| MFCC branch   | 22.3703% (5.2471%) | 43.7037% (7.7454%) | 21.7037% (5.3307%)  |
| GSV branch    | 56.7407% (4.3361%) | 89.8518% (2.9845%) | 98.2222% (1.1152%)  |
| Fusion branch | 52.1481% (2.0717%) | 86.4444% (4.3199%) | 63.5555% (13.1018%) |

in (Zeng et al., 2023), we also used 39-dimensional MFCC vectors. The features were extracted using 16 ms Hamming windows with an 8 ms overlap.

Because the audio samples were only 8 seconds long, the voice activity detection used in the non-speech approach did not produce enough frames for the GMM training. So, for the non-speech approach, residual noise signals were obtained by subtracting an estimated clean speech signal from the original speech signal.

To enable a fair comparison between the accuracy of our approaches to those disclosed in (Zeng et al., 2023), we emulated the results that would have been obtained on the small sample of the CCNU Mobile dataset by using the UBM trained on the TIMIT dataset of the original paper. In every experiment, the neural network training is done for 100 epochs using a batch size of 16. The learning rate is 0.001 and decreases by 1/10 every 20 epochs. The Adaptive Moment Estimator (Adam) optimizer is used. Brand and model identification accuracy using Zeng’s UBM are shown in Tables 11-12. The GMM hyperparameters were fixed for this experiment, so every classification was repeated 10 times and the mean and standard deviation were reported.

The MFCC branch of the network is the module that works with the MFCCs only, the GSV branch is the module that works with the GSVs only, and the fusion branch is the module that fuses the results of the previous two branches to obtain a final classification. Implementation details can be found in (Zeng et al., 2023). It is seen that without enough data, the GSV branch performs better than the MFCC branch and the fusion branch in both tasks.

The results of the traditional, the non-speech, and the UBM approach, when ML, SVM, and neural classification are applied, are summarized in Tables 13-

14. The UBM here is also trained on the TIMIT dataset. For the ML classification, 64 component GMMs are used. For the GSVs fed to the SVMs, 1 component is used, while for the SVMs applied to the UBM-adapted GSVs, 4 components are used.

In both tasks, the non-speech MFCCs improve the accuracy of all classification approaches compared to the traditional approach. This is also true when the UBM is used. Examining the neural network classification a little closer (i.e., the last 3 table rows in Tables 13 and 14), we see that the GSV branch using the UBM-adapted GSVs gives the best mean accuracy for both brand and model identification when the neural classification is employed. Zeng’s UBM (Tables 11-12) yields similar results to our UBM, as expected, except the fusion branch for model identification, where Zeng’s UBM attains a higher mean accuracy but a similar large standard deviation.

## 6 CONCLUSIONS

This paper has disclosed empirical evidence demonstrating that either the non-speech or UBM approaches increase mobile phone identification classification accuracy, even under data scarcity and noisy recordings. The non-speech approach performs better when a lot of speaker variability is observed since the features extracted from the non-speech segments of the signal are not corrupted by speaker-related noise. The UBM approach performs better in the presence of recording noise because of the probabilistic “matching” of the extracted features with the UBM’s components, which is robust to noisy features.

Considering future work, more experiments should be conducted on other datasets, especially the entire CCNU Mobile dataset if it becomes publicly

available, to check whether our results on the small sample are similar to the whole dataset. In the non-speech approach, instead of the MFCCs, which are designed based on the logarithmic scale of the human auditory system, other features providing the same resolution in all frequency ranges might also yield better results since the features are extracted from noise-like signals and not speech.

## ACKNOWLEDGEMENTS

This research was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the “2nd Call for H.F.R.I Research Projects to support Faculty Members & Researchers” (Project Number: 3888).

## REFERENCES

- Alsulaiman, M., Muhammad, G., Bencherif, M. A., Mahmood, A., and Ali, Z. (2013). Ksu rich arabic speech database. *Information (Japan)*, 16(6 B):4231–4253.
- Baldini, G. and Amerini, I. (2019). Smartphones identification through the built-in microphones with convolutional neural network. *IEEE Access*, 7:158685–158696.
- Baldini, G. and Amerini, I. (2022). Microphone identification based on spectral entropy with convolutional neural network. In *Proc. 2022 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE.
- Berdich, A., Groza, B., Levy, E., Shabtai, A., Elovici, Y., and Mayrhofer, R. (2022). Fingerprinting smartphones based on microphone characteristics from environment affected recordings. *IEEE Access*, 10:122399–122413.
- Berdich, A., Groza, B., and Mayrhofer, R. (2023). A survey on fingerprinting technologies for smartphones based on embedded transducers. *IEEE Internet of Things Journal*, 10(16):14646–14670.
- Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.
- Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (methodological)*, 39(1):1–22.
- Garcia-Romero, D. and Espy-Wilson, C. Y. (2010). Automatic acquisition device identification from speech recordings. In *Proc. 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1806–1809. IEEE.
- Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. (1988). Getting started with the darpa timit cd-rom: An acoustic phonetic continuous speech database. *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, 107:16.
- Giganti, A., Cuccovillo, L., Bestagini, P., Aichroth, P., and Tubaro, S. (2022). Speaker-independent microphone identification in noisy conditions. In *Proc. 2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1047–1051. IEEE.
- Hanilci, C., Ertas, F., Ertas, T., and Eskidere, Ö. (2011). Recognition of brand and models of cell-phones from recorded speech signals. *IEEE Transactions on Information Forensics and Security*, 7(2):625–634.
- Hanilçi, C. and Kinnunen, T. (2014). Source cell-phone recognition from recorded speech using non-speech segments. *Digital Signal Processing*, 35:75–85.
- Khan, M. K., Zakariah, M., Malik, H., and Choo, K.-K. R. (2018). A novel audio forensic data-set for digital multimedia forensics. *Australian Journal of Forensic Sciences*, 50(5):525–542.
- Kotropoulos, C. (2014). Source phone identification using sketches of features. *IET Biometrics*, 3(2):75–83.
- Kotropoulos, C. and Samaras, S. (2014). Mobile phone identification using recorded speech signals. In *Proc. 2014 19th International Conference on Digital Signal Processing*, pages 586–591. IEEE.
- Leonzio, D. U., Cuccovillo, L., Bestagini, P., Marcon, M., Aichroth, P., and Tubaro, S. (2023). Audio splicing detection and localization based on acquisition device traces. *IEEE Transactions on Information Forensics and Security*, 18:4157–4172.
- Maher, R. (2009). Audio forensic examination. *IEEE Signal Processing Magazine*, 26(2):84–94.
- Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, 9(5):504–512.
- Qamhan, M., Alotaibi, Y. A., and Selouani, S. A. (2023). Transformer for authenticating the source microphone in digital audio forensics. *Forensic Science International: Digital Investigation*, 45:301539.
- Reynolds, D., Quatieri, T., and Dunn, R. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41.
- Reynolds, D. A. (1997). HTIMIT and LLHDB: speech corpora for the study of handset transducer effects. In *Proc. 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1535–1538. IEEE.
- Sohn, J., Kim, N. S., and Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3.
- Zeng, C., Feng, S., Zhu, D., and Wang, Z. (2023). Source acquisition device identification from recorded audio based on spatiotemporal representation learning with multi-attention mechanisms. *Entropy*, 25(4):626.
- Zeng, C., Zhu, D., Wang, Z., Wang, Z., Zhao, N., and He, L. (2020). An end-to-end deep source recording device identification system for web media forensics. *International Journal of Web Information Systems*, 16(4):413–425.