Telephone Handset Identification by Collaborative Representations

Yannis Panagakis[1] and Constantine Kotropoulos

Department of Informatics
Aristotle University of Thessaloniki
email: {panagakis,costas}@ aiia.csd.auth.gr

Abstract

Recorded speech signals convey information not only for the speakers' identity and the spoken language, but also for the acquisition devices used for their recording. Therefore, it is reasonable to perform acquisition device identification by analyzing the recorded speech signal. To this end, recording-level spectral, cepstral, and fusion of spectral and cepstral features are employed as suitable representations for device identification. The feature vectors extracted from the training speech recordings are used to form overcomplete dictionaries for the devices. Each test feature vector is represented as a linear combination of all the dictionary columns (i.e., atoms). Since the dimensionality of the feature vectors is much smaller than the number of training speech recordings, there are infinitely many representations of each test feature vector with respect to the dictionary. These representations are referred to as collaborative representations in the sense that all the dictionary atoms collaboratively represent any test feature vector. By imposing the representation to be either sparse (i.e., to admit the minimum ℓ_1 norm) or to have the minimum ℓ_2 norm, unique collaborative representations are obtained. The classification is performed by assigning each test feature vector the device identity of the dictionary atoms yielding the minimum reconstruction error. This classification method is referred to as the sparse representation-based classifier (SRC) if the sparse collaborative representation is employed and as the least squares collaborative representation-based classifier (LSCRC) in the case of the minimum ℓ_2 norm regularized collaborative representation is used for

[1] Corresponding author: Yannis Panagakis, Department of Informatics, Aristotle University of Thessaloniki, Box 451, Thessaloniki, GR-54124, Greece, email: yannisp@csd.auth.gr

reconstructing the test sample. By employing the LSCRC, state of the art identification accuracy of 97.67% is obtained on a set of 8 telephone handsets, from Lincoln-Labs Handset Database.

Keywords: Digital Speech Forensics, Audio Forensics, Telephone Handset Identification, Collaborative Representation, Sparse Representation.

1 Introduction

Speech is the most natural way to communicate between humans. Nowadays, speech communication systems acquire, transmit, store, and process the information in digital form. However, the digital speech content can be imperceptibly altered by malicious, even amateur, users by using a variety of low-cost audio editing software. This creates a serious threat to the *knowledge life cycle*. Indeed, when hearing is no longer believing, the process of going from data to information, knowledge, understanding and, decision making is severely compromised (Farid, 2008). The consequences of this threat permeate a wide variety of fields, such as intellectual property, intelligence gathering, forensics, and news reporting to name a few. Currently, the methods to combat this threat in the field of *digital speech forensics* are still in their infancy. Therefore, there is an urgent need to advance the state-of-the-art in this field (Garcia-Romero & Espy-Wilson, 2010).

A first step to remedy the aforementioned threat is to extract forensic evidence about the mechanism involved in the generation of the speech recording by analyzing only the speech signal (Garcia-Romero & Espy-Wilson, 2010). That is, to identify the acquisition device by assuming that the devices along with their associated signal processing chain leave behind *intrinsic traces* in the speech signal. Indeed, the electronic devices, especially when including a microphone, cannot have exactly the same frequency response due to tolerances in the production of their electronic components and the different designs employed by the various manufacturers (Hanilci, 2012). 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 (Garcia-Romero & Espy-Wilson, 2010).

Although there are significant advances in image forensics (**Error! Reference source not found.**2008), audio forensics are less developed (**Error! Reference source not found.**). Few exceptions include the authentication of MP3 (Yang, 2008) and the authentication of speakers' environment (Oermann, 2005; Kraetzer, 2007; Malik & Farid, 2010). Similarly, a few automatic acquisition device identification systems have been developed. For instance, a method for the classification of 4 microphones has been proposed in (Kraetzer, 2007). The speech signal is parameterized by employing time domain features and the mel-frequency cepstral coefficients (MFCCs). The identification of the microphones is performed by a Naive Bayes classifier at a short-time frame level. Accuracies on the order of 60-75% have been reported. In (Garcia-Romero & Espy-Wilson, 2010), the identification of 8 landline telephone handsets and 8 microphones is addressed. In particular, the intrinsic characteristics of the device are captured by a template constructed by appending together the means of a Gaussian mixture which have been trained using linear and mel-scaled cepstral coefficients extracted by speech recordings of each device. Classification accuracies higher than 90% have been achieved, when a support vector machine (SVM) classifier was used. Recently, a robust system for the identification of cell-phones has been proposed in (Hanilci, 2012). In particular, when the MFCCs, extracted from speech recordings, are classified by an SVM, 14 different cell-phones are identified with an accuracy of 96.42%.

In this paper, a novel blind-passive method for landline telephone handset identification is proposed. The method resorts on spectral and cepstral based features extracted from speech recordings and their *collaborative representation* (Zhang, 2011), revealing the identity of the recording device. Fig. 1 outlines the proposed method. In particular, 5 recording-level features are used for device characterization. The *random spectral features* (RSFs) (Panagakis & Kotropoulos, 2012a) and the *labeled spectral features* LSFs (Panagakis & Kotropoulos, 2012b) are obtained by applying unsupervised and supervised feature selection on the mean spectrogram of each speech recording, respectively. That is, for unsupervised feature selection, the dimensionality of the mean spectrogram is reduced by random projections (Bingham & Mannila, 2001) yielding the RSFs (Panagakis & Kotropoulos, 2012a). In the supervised setting, the label information (i.e., the class where each device belongs to) of the training speech recordings is taken into account in order to derive a mapping between the feature space where the mean spectrograms lie onto and the label space. The mapping between the aforementioned two spaces is obtained by solving a regression problem

and it is exploited next in order to extract discriminant features by test mean spectrograms. These features are referred to as LSFs (**Error! Reference source not found.**b). Apart from the aforementioned spectral features, the MFCCs (i.e., cepstral-based features) are considered for acquisition device characterization. The MFCCs have been successfully employed in device identification (Garcia-Romero & Espy-Wilson, 2010; Hanilci, 2012), since the logarithm involved in their calculation has additive property in the spectrum magnitude domain and therefore they can be considered as a superposition of latent variables related to the recording device and the speech content. Furthermore, in order to combine the discriminative power of spectral and cepstral features in device identification, an augmented feature vector is constructed by concatenating the mean spectrogram and the mean MFCCs of each speech recording. By applying random projections onto the augmented feature vector, the *random fusion features* (RFFs) are derived. Moreover, in analogy to LSFs, the *labeled fusion features* (LFFs) are obtained by applying supervised future selection on the concatenation of spectral and cepstral features.

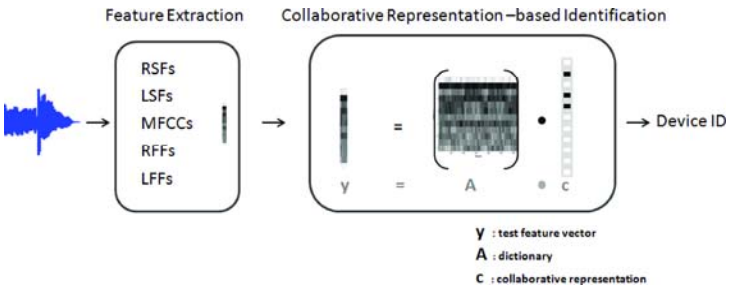

Figure 1: Each speech recording is represented by recording-level features. That is, the RSFs, the LSFs, the MFCCs, the RFFs, and the LFFs are extracted as features suitable for device identification. The features extracted from the training speech recordings are used to form overcomplete dictionaries (i.e., **A**) that collaboratively represent any feature vector **y,** extracted from a test speech recording. The collaborative representation **c**, is either a sparse solution of the linear system **Ac** = **y** (i.e., having the minimum $\| \mathbf{c} \|_1$) or a solution that admits the minimum ℓ_2 norm. It is used to identify the acquisition device of the recording.

The RSFs, the LSFs, the MFCCs, the RFFs, and the LFFs are used to form overcomplete dictionaries of basis atoms for devices' intrinsic traces. If sufficient training speech recordings are available for each device, it is possible to represent any vector of RSFs, LSFs, MFCCs, RFFs, or LFFs, extracted from an unknown (test) device, as a linear combination of all the dictionary atoms. These representations are

infinitely many, since the size of the feature vectors is much smaller than the number of the training speech recordings. They are referred to as collaborative representations in the sense that all the dictionary atoms (i.e., the training feature vectors) collaboratively represent any test feature vector (Chi & Porikli 2012; **Error! Reference source not found.**). If the collaborative representation is enforced to be sparse, that is to involve only a small fraction of the dictionary atoms, the *sparse representation-based classifier* (SRC) (**Error! Reference source not found.**) is derived. The SRC first encodes any test sample as a sparse linear combination of all the dictionary atoms via ℓ_1 norm optimization. The classification is performed by assigning each test feature vector the device identity (ID) of the dictionary atoms yielding the minimum reconstruction error. Although, the non-zero coefficients in sparse representation are associated with the dictionary atoms from the class that the test vector belongs to, **Error! Reference source not found.**(2011) indicate that it is the collaborative representation *and not* the ℓ_1 norm sparsity that makes the SRC a powerful classifier. Furthermore, Chi & Porikli (2012) demonstrate that, by enforcing the collaborative representation to have the minimum ℓ_2 norm, instead of the ℓ_1 norm, a similar classification accuracy can be achieved. The main advantage of the latter approach is the reduced computational complexity. Consequently, the *least squares collaborative representation-based classifier* (LSCRC) (Chi & Porikli, 2012) is an appealing alterative to the SRC. The aforementioned two constrained collaborative representations of spectral and cepstral features are employed for speech acquisition device identification.

The performance of the proposed methods in the identification of 8 telephone handsets is assessed by conducting experiments on the Lincoln-Labs Handset Database (LLHDB) (**Error! Reference source not found.**), when a stratified 2-fold cross-validation is applied. The collaborative representation-based classifiers (i.e., the SRC and the LSCRC) are compared against the linear SVM (**Error! Reference source not found.** & Lin, 2011) and the nearest-neighbor (NN) classifier, which employs the cosine similarity measure.

The experimental results demonstrate the effectiveness of the spectral features (i.e., the RSFs and LSFs) over the cepstral ones (i.e., the MFCCs) as representations capturing the device intrinsic characteristics, no matter which classifier is employed. Meanwhile, an accuracy of 97.67% in device identification is obtained, when the LFFs are classified

by the LSCRC, outperforming the state-of-the-art method (Garcia-Romero & Espy-Wilson, 2010) on the LLHDB dataset.

The paper is organized as follows. In Section 2, the calculation of the RSFs, the LSFs, the MFCCs, the RFFs, and the LFFs is described. The collaborative representation-based device identification is detailed in Section 3. The dataset and the experimental results are presented in Section 4. Conclusions are drawn in Section 5.

2 Features for Device Characterization

In this section, the calculation of spectral (i.e., the RSFs and the LSFs), cepstal features (i.e., the MFCCs) and the features derived by their fusion of them (i.e., the RFFs and LFFs) is described.

Spectral Features

The majority of features employed in tasks, such as speech and speaker recognition, spoken language identification, etc. are based on the spectrum of the speech signal. Assuming that the acquisition device is a linear time-invariant system, the impact of the acquisition device on the recorded speech can be modeled by the convolution of the original speech and the impulse response of the device. Thus, the identity of each acquisition device is embedded into the recorded speech, since the spectrum of any windowed recorded speech segment is the product of the spectrum of the original speech signal and the device frequency response.

Motivated by the aforementioned assumption, the RSFs and the LSFs are proposed here for device characterization. These features are derived by applying unsupervised and supervised feature selection to the mean spectrograms of the recordings, respectively. The spectrogram of each recorded speech signal is calculated by employing frames of duration 64 ms with a hop size of 32 ms and 2048 DFT bins. Then, the logarithm of the spectrogram is calculated and averaged along the time axis, yielding a 2048-dimensional mean spectrogram.

The RSFs are obtained as follows. The dimensionality of the mean spectrogram is reduced to d<2048 by employing a d×2048 orthogonal random Gaussian matrix $\mathbf{Q}$, as described in (**Error! Reference source not found.** & Mannila, 2012). Clearly, random projections can be interpreted as an unsupervised feature selection method, since a number d out of 2048 features is selected for acquisition device representation. Let $\mathbf{X} \in \mathbb{R}^{d \times J}$ be the data matrix that contains J vectors of RSFs of size d in its columns. The entries of $\mathbf{X}$ are further post-processed as follows:

Each row of $\mathbf{X}$ is normalized to the range [0,1] by subtracting from each matrix element the row minimum and then by dividing it with the difference between the row maximum and the row minimum.

In order to extract discriminant features from the mean spectrograms the label information of the devices that belong to the training set is taken into account. In particular, we aim to derive features that are highly dependent on the labels. Let $\mathbf{X}_t \in \mathbb{R}^{2048\times J}$ be the training data matrix, containing in its columns the 2048-dimensional mean spectrograms extracted from, J speech signals recorded by using acquisition devices from K classes. Let also $\mathbf{L} \in \{0,1\}^{K\times J}$ be the label indicator matrix, where the kth component of the jth column of $\mathbf{L}$, $\mathbf{l}_j$, is 1 if the jth device belongs to class $k \in \mathcal{K}$,

Features highly dependent on the labels can be obtained by seeking a linear mapping $\mathbf{M} \in \mathbb{R}^{K\times 2048}$, such that the space of the training mean spectrograms is mapped onto the label space, i.e., $\mathbf{L} = \mathbf{M}\mathbf{X}_t$. The aforementioned problem can be casted as a regression problem, since it involves the identification of the relationship between sets of dependent variables and independent ones. Although, a simple least squares regression could be employed to derive $\mathbf{M}$, it is well known that such an approach suffers from overfitting. To remedy this drawback of the least squares regression, $\mathbf{M}$ is found by solving the following ridge regression problem:

$$\text{argmin}_{\mathbf{M}} \| \mathbf{L} - \mathbf{M}\mathbf{X}_t \|_F^2 + \lambda_1 \| \mathbf{M} \|_F^2, \quad (1)$$

where λ_1 is a regularization parameter (e.g., the value λ_1 = 0.5 was used in the experiments) and $\| \cdot \|_F$ denotes the Frobenius norm. The unique closed form solution of (1) is

$$\mathbf{M} = \mathbf{L}\mathbf{X}_t^T (\mathbf{X}_t \mathbf{X}_t^T + \lambda_1 \mathbf{I})^{-1}. \quad (2)$$

$\mathbf{I}$ denotes the identity matrix of compatible dimensions. In the test phase, by premultiplying any mean spectrogram by $\mathbf{M}$ the K-dimensional vector of the LSFs is obtained. The overall procedure of spectral feature extraction is depicted in Fig. 2.

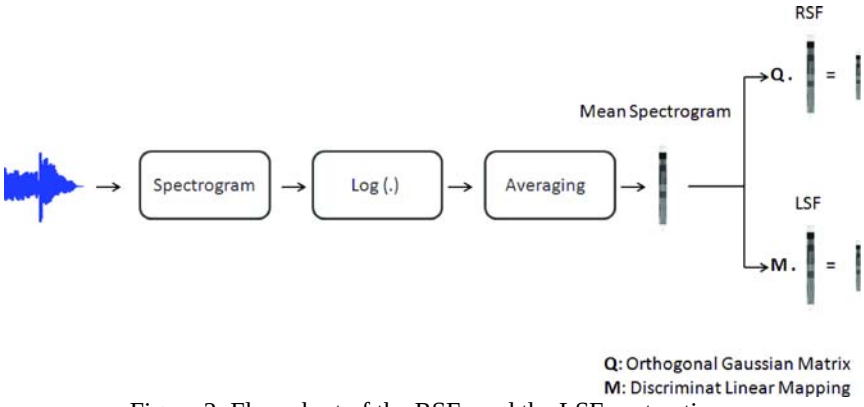

Figure 2: Flow chart of the RSFs and the LSFs extraction.

Cepstral Features

The MFCCs are considered as baseline features (Garcia-Romero & Espy-Wilson, 2010). They encode the frequency content of the speech signal by parameterizing the rough shape of spectral envelope. The success of the MFCCs in device identification is justified in (Hanilci, 2012). Roughly speaking, the logarithm, which involved in the calculation of the MFCCs is a nonlinear transformation with additive property in the spectrum magnitude domain and thus the cepstral features can be considered as a superposition of latent variables, which are related to the recording device, and variables which, are related to the speech content. Following (Garcia-Romero & Espy-Wilson, 2010), 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 along the log-energies of the bands. The sequence of 23-dimensional MFCCs is averaged along the time axis yielding a 23-dimensional mean vector.

In Figs. 3 and 4, the mean spectrograms and the MFCCs are depicted, for the same speech utterance recorded by 8 different telephone handsets, respectively. Clearly, both the mean spectrograms and the mean MFCCs convey discriminant information for the recording device.

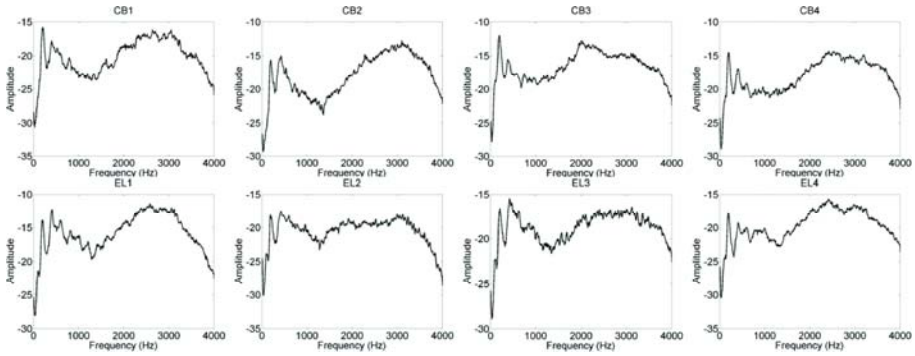

Figure 3: Mean spectrograms of a speech utterance recorded by 8 different telephone handsets in LLHDB.

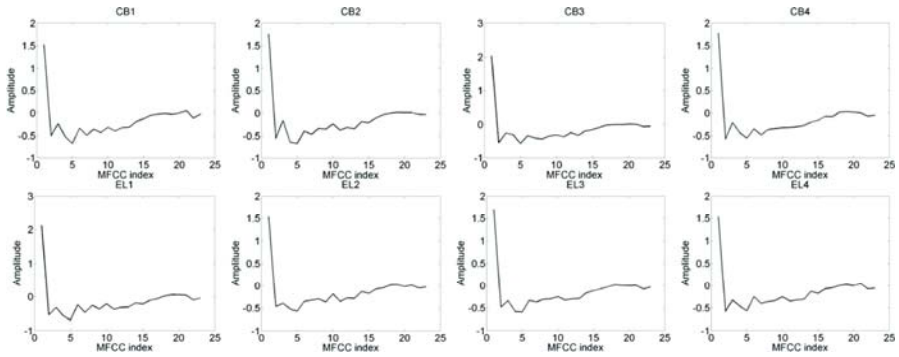

Figure 4: 23-dimensional mean MFCCs of a speech utterance recorded by 8 different telephone handsets in LLHDB.

Fusion of Spectral and Cepstral Features

To exploit the discriminative power of spectral and cepstral features in acquisition device identification, an augmented feature vector is constructed by concatenating the mean spectrogram and the mean MFCCs of each speech recording. By applying random projections onto the 2048+23=2071-dimensional augmented feature vector, the RFFs are obtained in a similar manner with the RSFs. Furthermore, the LFFs are derived by seeking a linear mapping $\mathbf{M} \in \mathbb{R}^{K \times 2071}$, such that the space spanned by the training augmented feature vectors is mapped onto the label space as described above in case of LSFs.

The data matrices containing the MFCCs and the RFFs are post-processed as described previously for the RSFs.

3 Acquisition Device Identification via Collaborative Representations

Given a number of labeled feature vectors from *N* acquisition devices, overcomplete dictionaries are formed. The problem of revealing the device identity of each feature vector (which can be either RSFs, LSFs, MFCCs, RFFs, or LFFs here) is addressed by seeking the sparse or the minimum ℓ_2 norm constrained collaborative representation of the test feature vector with respect to the dictionary, containing in its columns the training feature vectors. That is, the SRC (Wright, 2009) and the LSCRC (Chi & Porikli, 2012) are employed. In the following, the size *d* of the RSFs and the RFFs is varying according to the projection matrix employed while *d*=*K* for the LSFs and LFFs. The size of the mean MFCCs is 23 (i.e., *d*=23).

Let us denote by $\mathbf{A}_i = [\mathbf{a}_{i,1} | \mathbf{a}_{i,2} | \ldots | \mathbf{a}_{i,N_i}] \in \mathbb{R}^{d \times N_i}$ the dictionary that contains N_i either RSFs, LSFs, MFCCs, RFFs, or LFFs stemming from the *i*th device as column vectors (i.e., dictionary atoms). Next, let $\mathbf{A} = [\mathbf{A}_1 | \mathbf{A}_2 | \ldots | \mathbf{A}_K] \in \mathbb{R}^{d \times J}$ be an overcomplete dictionary formed by concatenating *J* feature vectors, which stem from *K* acquisition devices. A test feature vector $\mathbf{y} \in \mathbb{R}^d$ can be collaboratively represented by all the dictionary atoms of **A** as:

$$\mathbf{y} = \mathbf{A}\mathbf{c}, \tag{3}$$

where $\mathbf{c} \in \mathbb{R}^J$ is the collaborative representation of the test feature vector $\mathbf{y} \in \mathbb{R}^d$ by employing all training feature vectors as a dictionary. Clearly, if *d*<<*J* (i.e., **A** is overcomplete) there are infinitely many collaborative representations of **y** with respect to **A**. To restrict the solution space of the underdetermined linear system (3) to a single solution, one can impose norm constraints on the collaborative representation (i.e., **c**). Two popular constraints are: 1) the sparsity (i.e., only a few elements of **c** are non-zero) and 2) the ℓ_2 norm of **c** is minimal. Based on the aforementioned constraints, two powerful collaborative representation-based classifiers, namely the SRC and the LSCRC are derived.

Sparse Collaborative Representation

The problem of *sparse collaborative representation* is to find the coefficient vector **c**, such that $\mathbf{y} = \mathbf{A}\mathbf{c}$ and $\|\mathbf{c}\|_0$ is minimized, i.e.,

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} \| \mathbf{c} \|_0 \quad s.t. \ \mathbf{A}\mathbf{c} = \mathbf{y}, \tag{4}$$

where $\| \cdot \|_0$ is the ℓ_0 quasi-norm returning the number of the non-zero entries of a vector. Finding the solution of the optimization problem (4) is NP-hard due to the nature of the underlying combinatorial optimization. An approximate solution to the problem (4) can be obtained by replacing the ℓ_0 norm with the ℓ_1 norm:

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} \| \mathbf{c} \|_1 \quad s.t. \ \mathbf{A}\mathbf{c} = \mathbf{y}, \tag{5}$$

where $\| \cdot \|_1$ denotes the ℓ_1 norm of a vector. In (Donoho, 2006), it has been proved that if the solution is sparse enough, then the solution of (4) is equivalent to the solution of (5), which can be obtained by standard linear programming methods in polynomial time. A sparse collaborative representation aims to represents the test feature vector **y** using a minimal number of dictionary atoms. Furthermore, most of the non-zero entries of **c** correspond to atoms from the device class that **y** comes from.

Least Squares-Norm Collaborative Representation

By constraining the solution of (3) to has the minimum ℓ_2 norm, the problem of *least squares-norm collaborative representation* is:

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} \| \mathbf{c} \|_2 \quad s.t. \ \mathbf{A}\mathbf{c} = \mathbf{y}, \tag{6}$$

or equivalently:

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} \| \mathbf{c} \|_2 + \lambda_2 \| \mathbf{y} - \mathbf{A}\mathbf{c} \|_2^2, \tag{7}$$

where λ_2 is a regularization parameter and $\| \cdot \|_F$ denotes the ℓ_2 vector norm. The unique closed-form solution of (7) is obtained by:

$$\mathbf{c}^* = \left(\mathbf{A}^T\mathbf{A} + \lambda_2 \mathbf{I} \right)^{-1} \mathbf{A}^T\mathbf{y}. \tag{8}$$

Clearly, the computational cost of (8) is much smaller than that of solving (5), where iterative algorithms are employed.

Collaborative Representation-Based Classification

By solving either (5) or (7), a collaborative representation $\mathbf{c}^*$ of the test feature vector with respect to the corresponding dictionary is obtained and employed next for classification. Since the non-zero entries

in $\mathbf{c}^*$ are associated to multiple devices, each test feature vector is classified to the device class that minimizes the residual $r_i(\mathbf{y}) = \| \mathbf{y} - \mathbf{A}\delta_i(\mathbf{c}^*)\|_2^2 / \| \delta_i(\mathbf{c}^*)\|_2^2$, where $\delta_i(\mathbf{c}^*) \in \mathbb{R}^{N_i}$, $i = 1, 2, \ldots, 8$ is a new vector, whose nonzero entries are associated to the *i*th device only (Chi & Porikli, 2012; Wright, 2009). If the $\mathbf{c}^*$ is sparse, the aforementioned classifier is referred to as the SRC (**Error! Reference source not found.**), while if $\mathbf{c}^*$ is the solution of (7), the classifier is referred to as the LSCRC (Chi & Porikli, 2012).

It is worth mentioning that, the collaborative representation-based classifiers avoid under-fitting, since it employs multiple training samples for each class to linearly extrapolate the test sample, as opposed to the nearest one that the NN employs only. Furthermore, for each test sample, the number of samples needed is automatically determined. As a result, both the SRC and the LSCRC can better exploit the actual distributions of the training samples of each class. Therefore, they are likely to be more discriminant than other classifiers.

In Fig. 5 (a), the sparse collaborative representation coefficients $\mathbf{c}$ for a test vector of RSFs $\mathbf{y}$ extracted from a carbon-button telephone handset with the ID CB1 is isillustrated. Fig. 5 (b) shows the residual $r_i(\mathbf{y})$, computed by employing the sparse collaborative representation with respect to 8 telephone handset IDs. In Fig. 5 (c), the minimum ℓ_2 norm collaborative representation coefficients are depicted for the same test vector $\mathbf{y}$. Fig. 5 (d) shows the residual $r_i(\mathbf{y})$, *i*=1,2,…,8, computed by employing the minimum least-squares norm collaborative representation. It can be observed that, although the minimum least-squares norm collaborative representation is dense involing all the dictionary atoms in the representation of $\mathbf{y}$, the smallest residual corresponds to the correct telephone handset ID (i.e., CB1).

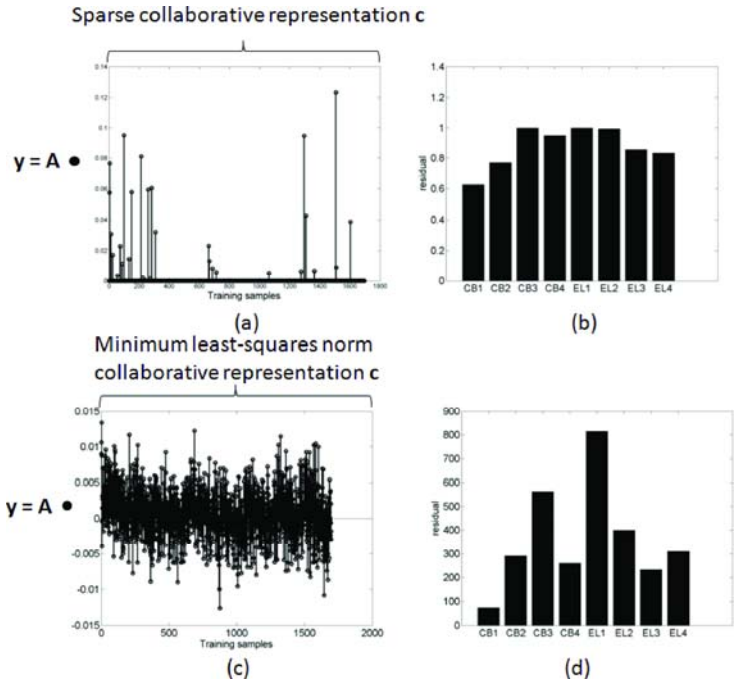

Figure 5: The test vector of RSFs **y** has been extracted by a carbon-button telephone handset with the ID: CB1. (a) The values of the sparse coefficients **c**. The non-zero entries of **c** are mainly associated with RSFs extracted from speech utterances recorded with the CB1. (b) The residuals $r_i(\mathbf{y})$, $i = 1, 2, \ldots, 8$ of the RSFs, computed by employing the sparse collaborative representation. (c) The values of the minimum least-squares norm coefficients $\mathbf{c}$. (d) The residuals $r_i(\mathbf{y})$, $i = 1, 2, \ldots, 8$, of the RSFs, computed based on the minimum least-squares norm collaborative representation. In both (b) and (d), the smallest residual value reveals the identity of the telephone handset (i.e., CB1).

4 Experimental Evaluation

In order to assess the performance of the proposed method in acquisition device identification, experiments were conducted on the same subset of the Lincoln-Labs Handset Database (LLHDB) (Reynolds, 2010) as in (Garcia-Romero & Espy-Wilson, 2010). This subset consists of speech recordings from 53 speakers (24 males and 29 females) acquired by 8 landline telephone handsets. The first 4 telephone handsets are are carbon-button (CB1-CB4) and the remaining 4 are electrect (EL1-EL4). Following the experimental set-up used in (Garcia-Romero & Espy-Wilson, 2010), stratified 2-fold cross-validation is employed in the experiments conducted on the LLHDB.

The best identification accuracies are summarized in Table 1, when the RSFs, the LSFs, the MFCCs, the RFFs, and the LFFs are classified by the LSCRC (Chi & Porikli, 2012) the SRC (Wright, 2009), the linear SVM (Chang & Lin, 2011), and the NN which employes the cosine similarity measure. By inspecting Table 1, it is clear that the RSFs, the LSFs, the RFFs, and the LFFs are able to identify the acquisition device committing less errors than the MFCCs, no matter which classifier is employed. Moreover, the LSFs and LFFs achieve state-of-the-art identification accuracy if they are fed to either the SVM, or the NN, or the SRC, or the LSCRC classifier. Clearly, the fusion of spectral and cepstral features yield robust representations (i.e., the RFFs and the LFFs) suitable for telephone handset identification. The LSCRC achieves the best reported identification accuracy (i.e., 97.67%) on the LLHDB.

Table 1: Best telephone handset identification accuracies achieved by the RSFs, the LSFs, the MFCCs, the RFFs, and the LFFs, when the LSCRC, the SRC, the linear SVM, and the NN are employed.

Features	**Feature dimension**	**Classifier**	**Accuracy (%)**
RSFs	850	LSCRC	94.63
RSFs	475	SRC	95.40
RSFs	850	SVM	94.72
RSFs	175	NN	87.64
LSFs	8	LSCRC	95.90
LSFs	8	SRC	97.14
LSFs	8	SVM	97.58
LSFs	8	NN	96.52
MFCCS	23	LSCRC	83.46
MFCCs	23	SRC	89.79
MFCCs	23	SVM	87.35
MFCCs	23	NN	81.95
RFFs	550	LSCRC	95.22
RFFs	550	SRC	95.90
RFFs	850	SVM	94.92
RFFs	475	NN	89.50
LFFs	8	LSCRC	97.67
LFFs	8	SRC	97.64
LFFs	8	SVM	97.64
LFFs	8	NN	97.46
MFCCs- based Gaussian supervector (Garcia-Romero & Espy-Wilson, 2010)		SVM	93.20

The performance of the RSFs and the RFFs in telephone handset identification as a function of features dimension (i.e., d) is depicted in

Fig. 6 (a) and Fig. 6 (b), respectively. It is clear that for d>200 the collaborative representation-based classifiers (i.e., the LSCRC and the SRC) outperforms the state-of-the-art reported in (Garcia-Romero & Espy-Wilson, 2010), demonstrating the robustness of the proposed approach in acquisition device identification. The accurate telephone handset identification by the RSFs and the RFFs and their collaborative representations is attributed to the following fact. It is well known that by projecting high-dimensional feature vectors (e.g., the mean spectrograms) onto an orthogonal random Gaussian matrix, the dictionary constructed by these features of reduced dimension, obeys the restricted isometry property (RIP) of a certain, appropriate order (Candes, 2005). When this property holds, it is implied that the test RSFs (or the RFFs), cannot be in the null space of **A** and thus they can be reconstructed as a linear combination of the columns of **A**. In contrast, the mean MFCCs have only 23 dimensions and to the best of our knowledge there is neither a theoretical proof nor experimental findings regarding the RIP of the dictionary constructed by employing the MFCCs.

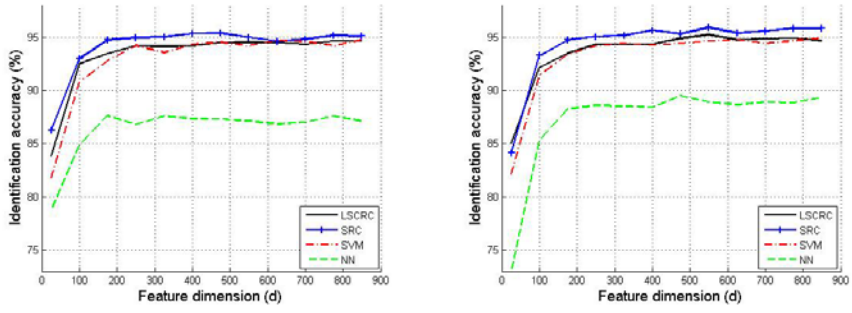

Figure 6: Telephone handsets identification accuracy for (a) the RSFs and (b) the RFFs obtained by the LSCRC, the SRC, the SVM, and the NN on the LLHDB.

The LSFs and the LFFs outperforms both the RSFs, the MFCCs, and the RFFs, since they are obtained following a supervised feature selection process. Moreover the LFFs yield higher accuracy in device identification than the LSFs. Insight to the performance achieved by LFFs when the LSCRC, the SRC, the SVM, and the NN classifier is

employed is offered by the confusion matrices shown in Fig. 7. The rows of the confusion matrices correspond to the predicted device and the columns indicate the actual device. The gray shading in these Figures highlights the fact that most of the identification errors remain within the transducer class (i.e., carbon-button and electrect). The carbon-button telephone handsets are identified more accurately than the electrect ones. This result is attributed to the fact that the transfer functions between the various carbon-button telephone handsets are quite different. Similar results were reported in (Garcia-Romero & Espy-Wilson, 2010).

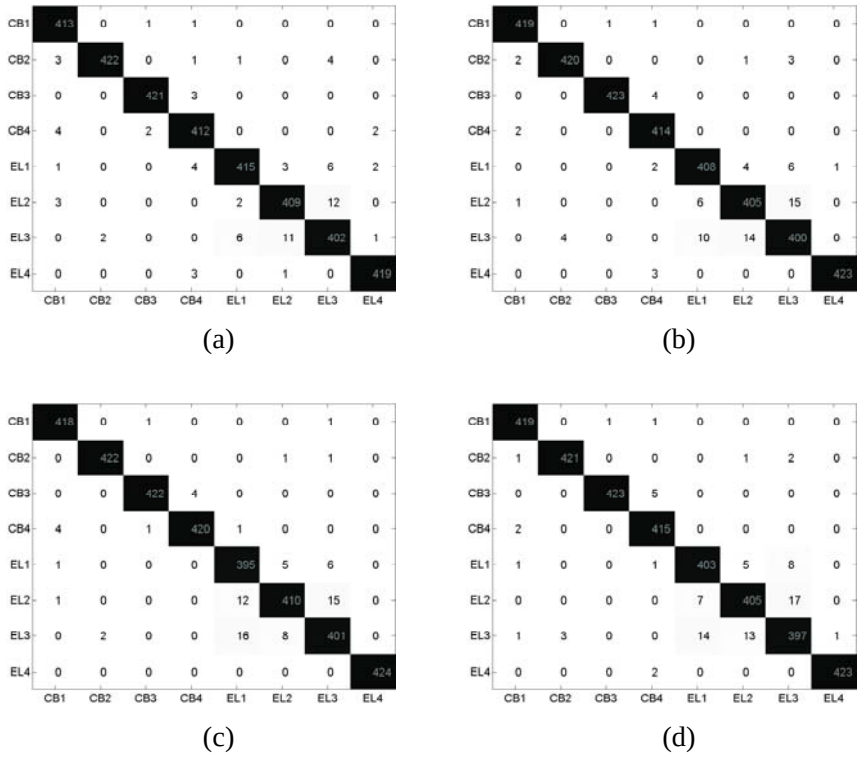

Figure 7: Confusion matrix for 8 telephone handsets based on the LFFs, when they are classified by (a) the LSCRC, (b) the SRC, (c) the SVM, and (d) the NN.

5 Conclusions

A promising method for telephone handset identification from speech signals has been proposed. Spectral and cepstral features have been demonstrated to capture the intrinsic trace of the acquisition device,

while the collaborative based classification has been shown to be able to identify the acquisition device. The experimental results indicate that the LFFs are the most suitable features for acquisition device characterization, yileding state of the art device identification accuracy when they are classified by the LSCR.

Acknowledgment

Y. Panagakis has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heraclitus II. Investing in Knowledge Society through the European Social Fund. C. Kotropoulos has been supported by the Cost Action IC 1106 "Integrating Biometrics and Forensics for the Digital Age".

References

Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: applications to image and text data. In *Proc. 7th ACM Int. Conf. Knowledge Discovery and Data Mining* (pp. 245-250). San Francisco, California, USA.

Candes, E., & Tao, T. (2005). Decoding by linear programming. *IEEE Trans. Inform. Theory (51), 4203-4215*

.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. (2), 1-27.*

Chi, Y., & Porikli, F. (2012). Connecting the dots in multi-class classification: From nearest subspace to collaborative representation, In *Proc. 2012 IEEE Conf. Computer Vision and Pattern Recognition* (pp. 3602-3609). Washington, DC, USA.

Donoho, D. (2006). For most large underdetermined systems of equations, the minimal l1-norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics (59), 907-934.*

Farid, H. (2008). Digital image forensics. *Scientific American (6), 66-71*.

Garcia-Romero, D., & Espy-Wilson, C. Y. (2010). Automatic acquisition device identification from speech recordings. *In Proc. 2010 IEEE Int.*

Conf. Acoustics, Speech, and Signal Processing (1806-1809). Dallas, Texas, USA.

Hanilci, C., Ertas, F., Ertas, T., & Eskidere, O. (2012). Recognition of brand and models of cell-phones from recorded speech signals. *IEEE Trans. Information Forensics and Security (7), 625-634.*

Kraetzer, C., Oermann, A., Dittmann, J., & Lang, A., (2007). Digital audio forensics: a first practical evaluation on microphone and environment classification. In: *Proc. 9th ACM Workshop Multimedia and Security* (pp. 63-74). Dallas, Texas, USA.

Maher, R., 2009. Audio forensic examination. *IEEE Signal Processing Magazine (26), 84–94*.

Malik, H., & Farid, H. (2010). Audio forensics from acoustic reverberation. In *Proc. 2010 IEEE Int. Conf. Acoustics Speech and Signal Processing* (pp. 1710-1713). Dallas, Texas, USA.

Oermann, A., Lang, A., & Dittmann, J. (2005). Verifier-tuple for audio-forensic to determine speaker environment. In *Proc. 7th ACM Workshop on Multimedia and Security* (pp. 57-62). New York, NY, USA.

Panagakis, Y., & Kotropoulos, C. (2012a). Automatic telephone handset identification by sparse representation of random spectral features. *In Proc. 2012 ACM Multimedia and Security* (pp. 91-96). Coventry, UK.

Panagakis, Y., & Kotropoulos, C. (2012b). Telephone handset identification by feature selection and sparse representations. In *Proc. 2012 IEEE Int. Workshop Information Forensics and Security* (pp. 73-78), Tenerife, Spain.

Reynolds, D. (1997). HTIMIT and LLHDB: speech corpora for the study of handset transducer effects. In *Proc. 1997 IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (pp. 1535-1538). Munich, Germany

Wright, J., Yang, A., Ganesh, A., Sastry, S., & Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell. (31), 210-227.*

Yang, R., Qu, Z., & Huang, J. (2008). Detecting digital audio forgeries by checking frame offsets. In *Proc. 10th ACM Workshop on Multimedia and Security* (pp. 21-26). New York, NY, USA.

Zhang, L., Yang, M., & Xiangchu, F. (2011). Sparse representation or collaborative representation: Which helps face recognition?. In: *Proc.*

2011 Int. Conf. Computer Vision (pp. 471-478), Washington, DC, USA.