Automatic Telephone Handset Identification by Sparse Representation of Random Spectral Features

Yannis Panagakis
Dept. of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
panagakis@aiia.csd.auth.gr

Constantine Kotropoulos
Dept. of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
costas@aiia.csd.auth.gr

ABSTRACT

Speech signals convey information not only for speakers' identity and the spoken language, but also for the acquisition devices used during their recording. Therefore, it is reasonable to perform acquisition device identification by analyzing the recorded speech signal. To this end, the *random spectral features* (RSFs) are proposed as an intrinsic fingerprint suitable for device identification. The RSFs are extracted from each speech signal by first averaging its spectrogram along the time axis and then by projecting the resulting mean spectrogram onto a Gaussian random matrix of compatible dimensions. By applying a sparse-representation based classifier to the device RSFs, state-of-the-art identification accuracy of 95.55% has been obtained on a set of 8 telephone handsets, from Lincoln-Labs Handset Database (LLHDB).

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications—*Signal Processing*

General Terms

Theory, Measurement, Reliability, Experimentation, Verification

Keywords

Digital speech forensics, random features, 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 [5]. 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 theories and tools 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 [6].

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 [6]. That is, to identify the acquisition device by assuming that the acquisition devices along with their associated signal processing chain leave behind *intrinsic traces* in the speech signal. Indeed, the electronic devices, especially when include a microphone, cannot have exactly the same frequency response due to tolerances in the production of the electronic components and the different designs employed by the various manufacturers [7]. 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 [6].

Although there are significant advances in image forensics [5], audio forensics are less developed [9]. Few exceptions include the authentication of MP3 [14] and the authentication of speakers' environment [11, 8, 10]. 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 [8]. 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 [6], 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 trained on the speech recordings of each device. To this end, linear- and mel-scaled cepstral coefficients were employed for speech signal representation. Classification accuracies higher than 90% have been achieved, when a support vector machine

MM&Sec'12, September 6–7, 2012, Coventry, United Kingdom.

(SVM) classifier was employed. Recently, a robust system for the identification of cell-phones has been proposed in [7]. In particular, when the MFCCs extracted from device 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 suitable feature extraction from speech recordings and their sparse representation, enabling to trace the recording device. In particular, the *random spectral features* (RSFs) are proposed as intrinsic features for tracing the recording device. The RSFs are obtained as follows: the spectrogram of each speech recording is computed and it is averaged next along the time axis, yielding the mean spectrogram. Then, the dimensionality of the mean spectrogram is reduced by random projections [1] yielding the RSFs of speech recording. These RSFs form an overcomplete dictionary of basis signals for devices' intrinsic traces, which is exploited for *sparse representation-based classification* (SRC) [13]. If sufficient training speech recordings are available for each device, it is possible to express any RSFs extracted from an unknown (test) device as a compact linear combination of the dictionary atoms for the device actually used for its recording. This representation is designed to be sparse, because it involves only a small fraction of the dictionary atoms and can be computed efficiently via ℓ_1-norm optimization. The classification is performed by assigning each vector of test RSFs the device identity (ID) the dictionary atoms weighted by non-zero coefficients are associated with.

The performance of the proposed method in the identification of 8 telephone handsets is assessed by conducting experiments on the Lincoln-Labs Handset Database (LLHDB) [12] when a stratified 2-fold cross-validation is applied. For comparison purposes, the mean 23-dimensional MFCC vector of each speech recording is considered as a baseline feature for device characterization. Performance comparisons are made against the linear SVM [3] and the nearest-neighbor (NN) classifier, which employs the cosine similarity measure. The experimental results demonstrate the effectiveness of the proposed RSFs over the MFCCs as device fingerprints, no matter which classifier is employed. Meanwhile, the proposed device identification method yields an accuracy of 95.55%, outperforming the state-of-the-art method [6] on the LLHDB dataset.

The paper is organized as follows. In Section 2, the RSFs are introduced and the calculation of the MFCCs is described. The sparse 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. INTRINSIC FINGERPRINT EXTRACTION

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 are proposed as intrinsic traces of recording devices. The RSFs are obtained as follows. 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 FFT bins. Then, the logarithm of the spectrogram is calculated and averaged along the time axis, yielding a 2048-dimensional mean spectrogram. The dimensionality of the mean spectrogram is reduced to $d < 2048$ by employing a $d \times 2048$ orthogonal random Gaussian matrix, as described in [1]. The resulting d-dimensional RSFs are used for acquisition device representation.

Let $\mathbf{X} \in \mathbb{R}^{d \times s}$ be the data matrix that contains s 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.

The MFCCs are considered as baseline features [6]. They encode the frequency content of the speech signal by parameterizing the rough shape of spectral envelope. Following [6], 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. The data matrix containing the MFCCs is postprocessed as described previously for the RSFs.

In Figs. 1 and 2, the RSFs and the MFCCs are depicted, for the same speech utterance recorded by 8 different telephone handsets, respectively. Clearly, both the RSFs and the MFCCs convey discriminant information for the recording device.

3. ACQUISITION DEVICE IDENTIFICATION VIA SPARSE REPRESENTATION

The problem of revealing the device identity of a vector of RSFs, given a number of labeled RSFs from N acquisition devices is addressed based on the SRC [13].

Let us denote by $\mathbf{A}_i = [\mathbf{a}_{i,1} | \mathbf{a}_{i,2} | \dots | \mathbf{a}_{i,n_i}] \in \mathbb{R}^{d \times n_i}$ the dictionary that contains n_i RSFs stemming from the ith device as column vectors (i.e., dictionary atoms). Given a vector of test RSFs $\mathbf{y} \in \mathbb{R}^d$ that comes from the ith device, we can assume that $\mathbf{y}$ is expressed as a linear combination of the atoms that are associated to the ith device, i.e.,

$$\mathbf{y} = \sum_{j=1}^{n_i} \mathbf{a}_{i,j}\, c_{i,j} = \mathbf{A}_i\, \mathbf{c}_i \tag{1}$$

where $c_{i,j} \in \mathbb{R}$ are coefficients, which form the coefficient vector $\mathbf{c}_i = [c_{i,1}, c_{i,2}, \dots, c_{i,n_i}]^T$.

Next, let $\mathbf{A} = [\mathbf{A}_1 | \mathbf{A}_2 | \dots | \mathbf{A}_N] \in \mathbb{R}^{d \times n}$ be an overcomplete dictionary formed by concatenating n RSFs, which stem from N acquisition devices. Thus, the linear representation of $\mathbf{y} \in \mathbb{R}^d$ in (1) can be equivalently rewritten as

$$\mathbf{y} = \mathbf{A}\, \mathbf{c} \tag{2}$$

where $\mathbf{c} = [\mathbf{0}^T | \dots | \mathbf{0}^T | \mathbf{c}_i^T | \mathbf{0}^T | \dots | \mathbf{0}^T]^T$ is the $n \times 1$ augmented coefficient vector, whose elements are zero except

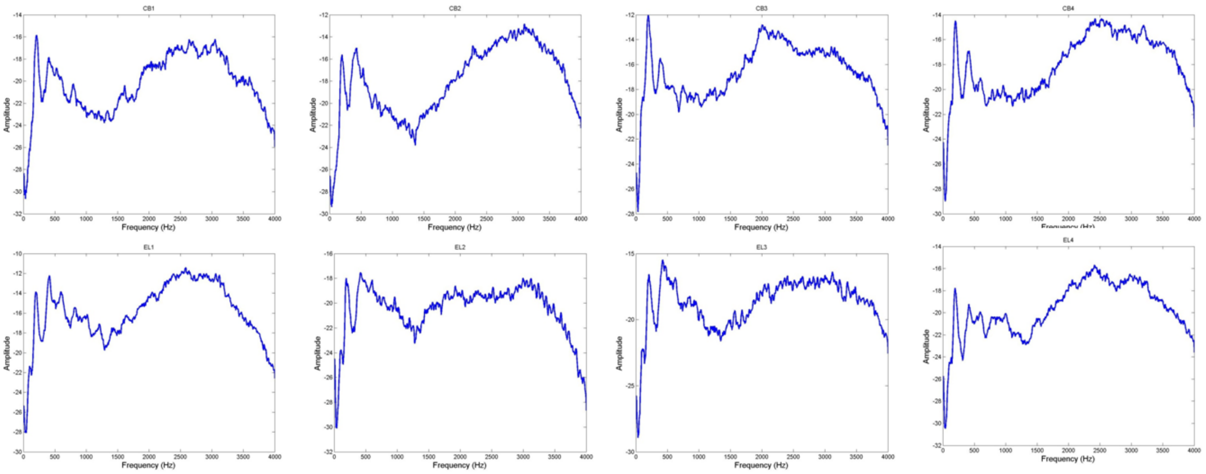


Figure 1: RSFs of a speech utterance recorded by 8 different telephone handsets in LLHDB.

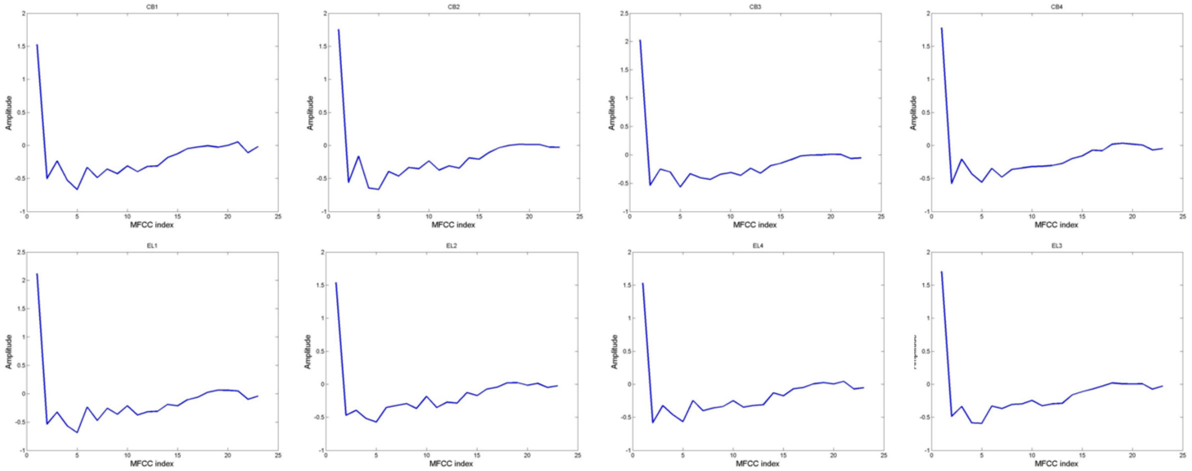


Figure 2: 23-dimensional mean MFCCs of a speech utterance recorded by 8 different telephone handsets in LLHDB.

those associated with the ith device. Thus, the entries of $\mathbf{c}$ contain information about the device the test vector of RSFs $\mathbf{y} \in \mathbb{R}^d$ comes from.

Since the device ID of a test vector of RSFs is unknown, we can predict it by seeking the sparsest solution to the linear system of equations $\mathbf{y} = \mathbf{A}\,\mathbf{c}$. Formally, given the overcomplete dictionary $\mathbf{A}$ and the vector of test RSFs $\mathbf{y} \in \mathbb{R}^d$, the problem of sparse representation is to find the coefficient vector $\mathbf{c}$, such that $\mathbf{y} = \mathbf{A}\,\mathbf{c}$ and $\|\mathbf{c}\|_0$ is minimized, i.e.,

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} \|\mathbf{c}\|_0 \quad \text{subject to } \mathbf{A}\mathbf{c} = \mathbf{y} \qquad (3)$$

where $\|.\|_0$ is the ℓ_0 quasi-norm returning the number of the non-zero entries of a vector. Finding the solution of the optimization problem (3) is NP-hard due to the nature of the underlying combinational optimization. An approximate solution to the problem (3) can be obtained by replacing the ℓ_0 norm with the ℓ_1 norm:

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} \|\mathbf{c}\|_1 \quad \text{subject to } \mathbf{A}\,\mathbf{c} = \mathbf{y} \qquad (4)$$

where $\|.\|_1$ denotes the ℓ_1 norm of a vector. In [4], it has been proved that if the solution is sparse enough, then the solution of (3) is equivalent to the solution of (4), which can be obtained by standard linear programming methods in polynomial time.

A test vector of RSFs can be classified as follows. The coefficient vector $\mathbf{c}^*$ is obtained by solving (4). Ideally, $\mathbf{c}^*$ contains non-zero entries in positions associated with the dictionary atoms (i.e., columns of $\mathbf{A}$) stemming from a single device, so that we can easily assign the vector of test RSFs $\mathbf{y}$ to that device. However, due to modeling errors, there are small non-zero entries in $\mathbf{c}^*$ that are associated to multiple devices. To cope with this problem, each RSF is classified to the device class that minimizes the residual $r_i(\mathbf{y}) = \|\mathbf{y} - \mathbf{A}\,\delta_i(\mathbf{c})\|_2$, where $\delta_i(\mathbf{c}) \in \mathbb{R}^n$ is a new vector, whose nonzero entries are associated to the ith device only [13].

In Fig. 3 (a), the sparse representation coefficients $\mathbf{c}$ for a test RSF vector $\mathbf{y}$ extracted from a carbon-button telephone handset with the ID CB1 are illustrated. Fig. 3 (b) shows the residual $r_i(\mathbf{y})$ with respect to 8 telephone handsets IDs.

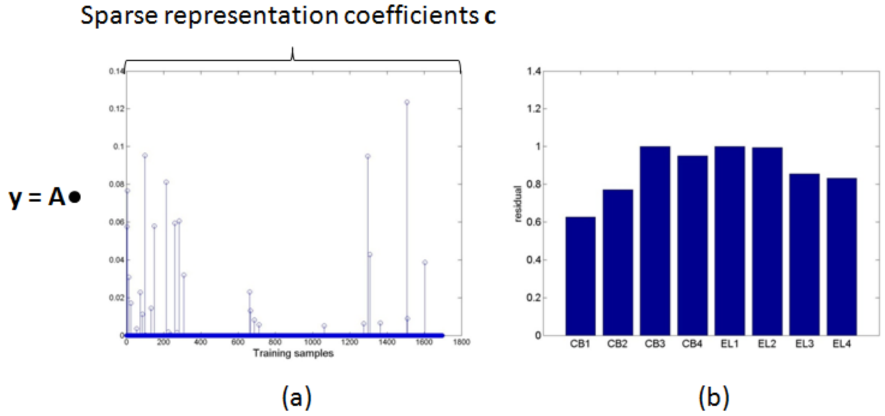


Figure 3: The test vector of RSFs y has been extracted by a carbon-button telephone handset with the ID: CB1. (a) The values of the sparse coefficients c. The non-zero entries of c are mainly associated with RSFs extracted from speech utterances recorded with the CB1. (b) The residuals $r_i(\mathbf{y})$ of the RSFs. The smallest residual value reveals the identity of the telephone handset (i.e., CB1).

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

Features	Feature dimension	Classifier	Accuracy (%)
RSFs	325	SRC	**95.55**
RSFs	625	SVM	94.81
RSFs	475	NN	88.23
MFCCs	23	SRC	89.79
MFCCs	23	SVM	87.35
MFCCs	23	NN	81.95
MFCCs- based Gaussian supervector[6]	N/A	SVM	93.20

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) [12] as in [6]. This subset consists of speech recordings from 53 speakers (24 males and 29 females) acquired by 8 landline telephone handsets. 4 of telephone handsets are carbon-button (CB1-CB4) and the remaining 4 are electrect (EL1-EL4). Following the experimental set-up used in [6], 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 and the MFCCs are classified by the SRC [13], the linear SVM [3], and the NN with the cosine similarity measure. By inspecting Table 1, it is clear that the RSFs are able to identify the acquisition device committing less errors than the MFFCs, no matter which classifier is employed. Moreover, the RSFs achieve state-of-the-art identification accuracy if they are fed to either the SVM or the SRC classifier. The latter classifier achieves the best reported identification accuracy (i.e., 95.55%) on the LLHDB. Similarly, the SRC outperforms the SVM, when the MFCCs are employed for device characterization.

The performance of the RSFs in telephone handset identification as a function of features dimension (i.e., d) is depicted in Fig. 4. It is clear that for $d > 200$ the SRC outperforms the best result reported in [6], demonstrating the robustness of the proposed approach in acquisition device identification.

Insight to the performance of the SRC is offered by the confusion matrix shown in Fig. 5 for $d = 325$. The gray shading in Fig. 5 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 [6].

The accurate telephone handset identification by RSFs and their sparse representations is attributed to the following fact. It is well known that by projecting the data onto an orthogonal random Gaussian matrix, the dictionary $\mathbf{A}$ obeys the restricted isometry property (RIP) of a certain, appropriate order (say S) [2]. When this property holds, $\mathbf{A}$ approximately preserves the Euclidean length of S-sparse RSFs, which in turn implies that S-sparse vectors cannot be in the null space of $\mathbf{A}$. The latter is needed since otherwise there would be no hope of reconstructing these vectors. Clearly, it cannot be guaranteed that the RIP holds for the dictionary constructed by employing the MFCCs as atoms.

5. CONCLUSIONS

A promising method for telephone handset identification from speech signals has been proposed. The RSFs have been

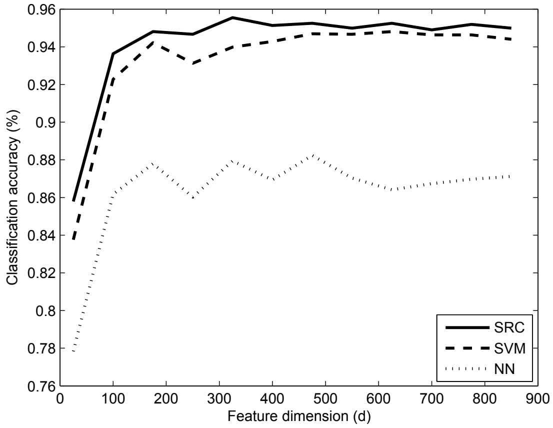
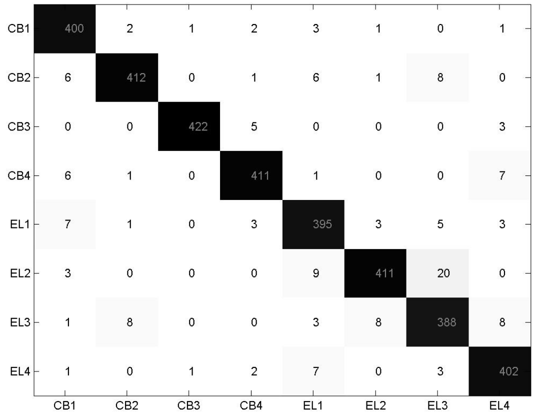


Figure 4: Telephone handsets identification accuracy for the RSF obtained by the SRC, the SVM, and the NN on the LLHDB.



	CB1	CB2	CB3	CB4	EL1	EL2	EL3	EL4
CB1	400	2	1	2	3	1	0	1
CB2	6	412	0	1	6	1	8	0
CB3	0	0	422	5	0	0	0	3
CB4	6	1	0	411	1	0	0	7
EL1	7	1	0	3	395	3	5	3
EL2	3	0	0	0	9	411	20	0
EL3	1	8	0	0	3	8	388	8
EL4	1	0	1	2	7	0	3	402

Figure 5: Confusion matrix for 8 telephone handsets based on the RSFs for $d = 325$ and their sparse representation on the LLHDB. The rows of the confusion matrix correspond to the predicted device and the columns indicate the actual device.

demonstrated to capture the intrinsic trace of the acquisition device, while the sparse representation-based classification has been shown to be able to identify the acquisition device. The experimental results validate the robustness of the RSFs over the MFCCs for device characterization, yielding a state-of-the-art performance in recognizing 8 telephone handsets from the LLHDB.

Acknowledgments

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".

6. REFERENCES

[1] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proc. 7th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pages 245–250, San Francisco, California, USA, 2001.

[2] E. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):4203–4215, 2005.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):1–27, 2011.

[4] D. Donoho. 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(7):907–934, 2006.

[5] H. Farid. Digital image forensics. *Scientific American*, 6(298):66–71, 2008.

[6] D. Garcia-Romero and C. Y. Espy-Wilson. Automatic acquisition device identification from speech recordings. In *Proc. 2010 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pages 1806–1809, Dallas, Texas, USA, 2010.

[7] C. Hanilci, F. Ertas, T. Ertas, and O. Eskidere. Recognition of brand and models of cell-phones from recorded speech signals. *IEEE Trans. Information Forensics and Security*, 7(2):625–634, 2012.

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

[9] R. Maher. Audio forensic examination. *IEEE Signal Processing Magazine*, 26(2):84–94, 2009.

[10] H. Malik and H. Farid. Audio forensics from acoustic reverberation. In *Proc. 2010 IEEE Int. Conf. Acoustics Speech and Signal Processing*, pages 1710–1713, Dallas, Texas, USA, 2010.

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

[12] D. Reynolds. HTIMIT and LLHDB: speech corpora for the study of handset transducer effects. In *Proc. 1997 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, volume 2, pages 1535–1538, Munich, Germany, 1997.

[13] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):210–227, 2009.

[14] R. Yang, Z. Qu, and J. Huang. Detecting digital audio forgeries by checking frame offsets. In *Proc. 10th ACM workshop on Multimedia and Security*, pages 21–26, New York, NY, USA, 2008.