Source Phone Identification Using Sketches of Features

Constantine L. Kotropoulos

Abstract

Speech recordings carry useful information for the devices used to capture them. Here, acquisition device identification is studied using *sketches of features* as intrinsic device characteristics. That is, starting from large-size raw feature vectors obtained by either averaging the log-spectrogram of a speech recording along the time axis or stacking the parameters of each component for a Gaussian Mixture Model modeling the speech recorded by a specific device, features of reduced size are extracted by mapping these raw feature vectors into a low-dimensional space. The mapping preserves the "distance properties" of the raw feature vectors. It is obtained by taking the inner product of the raw feature vector with a vector of independent identically distributed random variables drawn from a p-stable distribution. State-of-the art classifiers, such as a sparse representation-based classifier or support vector machines, applied to the sketches yield an identification accuracy exceeding 94% on a set of 8 landline telephone handsets from Lincoln-Labs Handset Database. Perfect identification is reported for a set of 21 cell-phones of various models from 7 different brands.

Index Terms

Digital speech forensics, symmetric p-stable distributions, sketches of features, sparse representations.

I. Introduction

Digital speech content can be imperceptibly altered by malicious, even amateur, users, employing a variety of low-cost audio editing software. This creates a serious threat permeating a wide variety of

Constantine Kotropoulos is with the Department of Informatics, Aristotle University of Thessaloniki, Box 451 Thessaloniki 541 24, GREECE.

Corresponding author: C. Kotropoulos, e-mail: costas@aiia.csd.auth.gr.

fields, such as intellectual property, intelligence gathering and forensics, to name a few [1]. Theories and tools to combat this threat in the field of *digital speech forensics* are still in their infancy [2].

First of all, one needs to extract forensic evidence about the mechanism involved in the generation of the speech recording by analyzing the speech signal [2]. That is, to identify the acquisition device by assuming that the device along with its associated signal processing chain leaves behind *intrinsic traces* in the speech signal. Indeed, the various devices (e.g., telephone handsets, cell-phones) do not have exactly the same frequency response due to the tolerance in the nominal values of the electronic components and the different designs employed by the various manufacturers [3]. 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 [2].

Audio forensics are less developed [4] than image forensics [1]. Several problems have attracted the interest of the forensics community, including codec identification, authentication of speakers' environment, and automatic acquisition device identification, so far. Many studies were performed for the identification of codecs, such as MP3 [5], Windows Media Audio codec [6], Code Excited Linear Prediction codecs [7], or G.711, G.726, G.728, G.729, Internet Low-Bit codec, Adaptive Multi-Rate NarrowBand, and Silk [8]. Classification and regression trees were reported to achieve an identification accuracy of 92% among nine codecs using a 50% cross-validation on a database with 180 test conditions, comprising three noise types (car, babble and hum) at five signal to noise ratios [9]. The authentication of speakers' environment was investigated in [10]–[13]. The effectiveness of Hidden Markov Model-based phone recognition[1] for forensic voice comparison has been evaluated in terms of both validity (accuracy) and reliability (precision) in [14]. Acoustic environment identification finds many applications (e.g., audio recording integrity authentication, real-time crime localization/identification). Statistical techniques for estimating the reverberation and background noise were proposed in [15], [16].

Telephone handset identification was first treated in order to avoid performance degradation in speaker recognition due to mismatches between training and test data. For example, autoassociative neural networks were reported to achieve an accuracy of 85% in a two-class problem (i.e., carbon-button vs. electret telephone handset identification) in the NIST-99 speaker evaluation database, employing 1448 test utterances [17]. A Gaussian mixture model-based handset selector was proposed in [18] and then

[1]In this sentence only, a phone refers to the acoustic realization of a phoneme in loose terms.

handset-specific stochastic second-order feature transformations were applied to the distorted feature vectors increasing speaker verification accuracy. Another method for the classification of 4 microphones was originally proposed in [11] and further improved thanks to a proper fusion strategy [12]. The speech signal was parameterized by employing time domain features and the mel-frequency cepstral coefficients (MFCCs). The identification of the microphones was performed by the Naive Bayes classifier at a short-time frame level. Accuracies in the order of 60–75% were reported. Rank level fusion was shown to increase the classification accuracy to 100% [12]. The identification of 8 landline telephone handsets and 8 microphones was addressed in [2]. In particular, the intrinsic characteristics of the device were captured by concatenating the mean vectors of a Gaussian mixture model (GMM) trained on the speech recordings of each device. Linear- and mel-scaled cepstral coefficients were employed for speech signal representation. A classification accuracy of 93.2% was reported for 8 landline telephone handset identification in the Lincoln-Labs Handset Database (LLHDB) [19], when a support vector machine (SVM) classifier and a 2-fold cross-validation was employed. The identification of 14 cell-phones was proposed in [3] extracting the MFCCs from each device speech recordings, which were then classified by an SVM. An identification accuracy of 96.42% was reported for 14 different cell-phones using a set of 3360 utterances uttered by 24 speakers equally divided into a training and test set. Blind-passive methods for landline telephone handset identification were proposed in [20] and [21]. More specifically, the random spectral features were extracted by reducing the size of average log-spectrograms thanks to an orthogonal random Gaussian projection matrix [20]. In a supervised setting, the label information (i.e., the class where each device belongs to) of the training speech recordings was taken into account in order to derive a mapping between the feature space where the average log-spectrograms lie onto and the label space [21]. This supervised method reached an accuracy of 97.58% in the LLHDB. The blind-passive method for landline telephone handset identification introduced in [20] was extended by investigating the *sketches of spectral features* (SSFs) as intrinsic traces suitable for device identification in [22]. The SSFs were extracted by taking the inner product of the average log-spectrogram with a vector of independent identically distributed (i.i.d.) random variables (r.vs) drawn from a p-stable distribution [23].

In this paper, the sketches (i.e., feature vectors of reduced size) are elaborated in a more wide sense than [22]. On the top of the SSFs, sketches are extracted from the Gaussian supervectors (GSVs). The GSVs are made by concatenating the model parameters of the GMM components (i.e., the mean vectors and/or the vectors comprising the elements of diagonal covariance matrices on their main diagonal) modeling the MFCCs, having excluded the 1st coefficient [24]. GSVs are extracted with or without resorting to a GMM universal background model (UBM) [25]. In the former case, the GSVs are made by

concatenating the mean vectors of the MFCCs after applying maximum a posteriori adaptation (MAP). The sketches of features form an overcomplete dictionary for devices' intrinsic traces. This dictionary is exploited then for *sparse representation-based classification* (SRC) [26]. If sufficient training speech recordings are available for each device, it is possible to express the sketches extracted from a recording captured by an unknown (test) device as a compact linear combination of the dictionary atoms for the device actually used during acquisition. 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 test sketch the device identity the dictionary atoms weighted by non-zero coefficients are associated with. It is demonstrated that by employing a proper p-stable distribution to randomly project feature vectors of large size to sketches of reduced size, very high recognition rates can be obtained by using either an SRC or an SVM.

The proposed method is tested for source phone identification in two databases. First, the SSFs are employed to identify the 8 landline telephone handsets of the LLHDB using a 2-fold cross-validation. That is, by dividing the available recordings into two disjoint sets balanced in the number of recordings, the number of speakers and their gender) as in [2], [3]. For comparison purposes, the average 23-dimensional MFCC vector of each speech recording is considered as a baseline feature for device characterization. Performance comparisons are made for the SRC, the linear SVM [27], [28], and the nearest-neighbor (NN) classifier, which employs the cosine similarity measure. The experimental results demonstrate the effectiveness of the SSFs over the MFCCs as device intrinsic traces, no matter which classifier is employed. The SSFs classified by the SRC yield an accuracy exceeding 95% on the LLHDB. The power of SSFs in landline telephone handset identification is also assessed, when the test speaker identity is left out from the training set. As expected, the device identification accuracy is reduced, but it still exceeds 78%. In the latter case, sketches of the GSVs are also extracted by MAP adapting the mean vectors of a GMM-UBM with 128 or 256 components trained on the training subset of the TIMIT database [29] with the mean vectors of the MFCCs extracted from the LLHDB utterances. The device identification accuracy increases to 94.11%, when the sketches of the GSVs are classified by the SVM.

Second, a database of 21 cell-phones of various models from 7 different brands was collected by recording 10 utterances uttered by 12 male speakers and another 12 female speakers, randomly chosen from the TIMIT database. The 10 utterances per speaker were concatenated in a single 30 s long recording. Two disjoint subsets were created, balanced in the number of files as well as speakers and gender. Each subset comprises 252 files. This database is referred to as the *MOBIPHONE* database, hereafter. A perfect device identification is achieved, when the sketches of the GSVs, extracted by MAP adapting

the mean vectors of a GMM-UBM with either 64 or 128 components trained on the training subset of the TIMIT database with the mean vectors of the MFCCs extracted from the MOBIPHONE recordings, are classified by any of the aforementioned classifiers. By omitting GMM-UBM training, GSVs are calculated by concatenating the mean vectors and the parameters of the GMM components (i.e., the mean vectors and/or the vectors comprising the elements of diagonal covariance matrices on their main diagonal) modeling the MFCCs extracted from the MOBIPHONE recordings. Using one component only, the sketches of these GSVs achieve an accuracy of 99.21%, when they are classified by the SRC. The device identification accuracy of SSFs classified by the SRC is 98.81%, while that of the average 23-dimensional MFCC vectors classified by the SVM is 97.22%.

The paper is organized as follows. In Section 2, the acquisition device intrinsic traces are introduced. That is, the calculation of the SSFs, the MFCCs, and the GSVs is described. The sparse representation-based device identification is detailed in Section 3. The LLHDB and the MOBIPHONE database as well as the experiments conducted are discussed in Section 4. Conclusions are drawn in Section 5.

II. Acquisition Device Intrinsic Traces

The majority of features employed in speech and speaker recognition, spoken language identification, or audio processing parameterize the spectrum of the signal. Assuming that the acquisition device is a linear time-invariant system, its impact on the recorded speech is modeled by the convolution of its impulse response and the original speech. Thus, the identity of the acquisition device is embedded into the recorded speech, since the spectrum of any recorded speech segment is the product of the spectrum of the original speech signal and the device frequency response.

A. Raw feature vectors

The first feature vector of large size is obtained from the spectrogram (i.e., the magnitude of the short-term Fourier transform) of each recorded speech signal by employing frames of duration 64 ms with a hop size of 32 ms and discrete Fourier transform of size 4096 samples. The speech frames are obtained by multiplying the recorded speech signal with a Hamming window. The resulting representation has a size of 2049 samples, including the dc value and 2048 equally spaced frequency samples of the (short-term) discrete-time Fourier transform in the interval $[0, f_s/2]$ for sampling frequency $f_s = 8$ KHz. Next, the logarithm of the spectrogram is calculated and is averaged along the time axis. The just described feature vector is called *average log-spectrogram*. The average log-spectrograms of the utterance SA1 “She had your dark suit in greasy wash water all years” from TIMIT database that was uttered by person FALW

and recorded by 4 carbon button and 4 electret landline telephone handsets in the LLHDB are depicted in Fig. 1 and 2, respectively.

The MFCCs are considered as the baseline features [2]. They encode the frequency content of the speech signal by parameterizing the rough shape of its spectral envelope. Following [2], the MFCC calculation employs frames of duration 20 ms with a hop size of 10 ms, and a 42-band filter bank. The correlation between the frequency bands is reduced by applying the discrete cosine transform to the log-energies of the bands. The sequence of 23-dimensional MFCCs is averaged along the time axis yielding a 23-dimensional average vector. In Fig. 3 and 4, the MFCCs are depicted, for the SA1 speech utterance uttered by the speaker FALW and recorded using the 4 carbon button and the 4 electret landline telephone handsets in the LLHDB database. Both the average log-spectrograms and the average MFCCs look appropriate as acquisition device intrinsic traces.

A second class of feature vectors having large size comprises the GSVs [25]. The GSVs are derived by either resorting to a GMM-UBM or not. In the former case, the mean vectors of the GMM-UBM are MAP adapted using the MFCCs of the source phone speech recordings. To do so, a GMM-UBM with a number of components ranging from 64 to 256 components was trained on the MFCCs extracted from the utterances of the training subset of the TIMIT database. For example, the GSV made by stacking the mean vectors of 128 components has size of 2816 elements, after having excluded the 1st MFCC coefficient. Let $\mathbf{m}_k$, $k = 1, \ldots, K$ denote the mean vectors of the GMM-UBM. The MAP adaptation of the mean vectors results in [25]

$$\boldsymbol{\nu}_k = \eta_k \, \mathbf{m}_k + (1 - \eta_k) \, \boldsymbol{\mu}_k \tag{1}$$

where $\boldsymbol{\mu}_k$ is the mean vector of the source phone MFCCs assigned to the kth component and η_k is a mixing coefficient. If $\zeta_{lk} = \mathrm{Prob}(k|\boldsymbol{\xi}_l)$ is the probability of assignment of the lth MFCC vector $\boldsymbol{\xi}_l$ of a source phone speech recording to the kth component, $\boldsymbol{\mu}_k$ in (1) is given by

$$\boldsymbol{\mu}_k = \frac{1}{\zeta_k} \sum_l \zeta_{lk} \, \boldsymbol{\xi}_l \tag{2}$$

where $\zeta_k = \sum_l \zeta_{lk}$. The mixing coefficient is given by

$$\eta_k = \frac{\zeta_k}{\zeta_k + \tau} \tag{3}$$

where τ is the so-called relevance factor, regulating the trade-off between what the data suggest and our prior belief contained in the UBM mean vectors. The scalar τ was set to 16 as recommended in [25]. Alternatively, one may choose to estimate the GMM model parameters directly from the MFCCs

extracted from the source phone recordings. In this case, GSVs of much smaller size are obtained by concatenating the parameters of the GMM components. For 5 components, the size of the resulting GSV is 220.

B. Sketches of features

Sketches, that is feature vectors of reduced size, are obtained by proper random projections. Denote the data matrix by $\mathbf{Z} \in \mathbb{R}^{d' \times n}$, containing the raw feature vectors of large size for the n recordings. For example, $d' = 2049$ for the average log-spectrograms, while $d' = 2816$ for the GVS made by concatenating the mean vectors for GMMs of 128 components. The size of such raw feature vectors is reduced to $d < d'$ by pre-multiplying $\mathbf{Z}$ with a projection matrix $\mathbf{R} \in \mathbb{R}^{d \times d'}$, yielding $\mathbf{X} = \mathbf{R}\,\mathbf{Z}$. The columns of $\mathbf{X}$ are the sketches proposed as acquisition device intrinsic traces. If the columns of $\mathbf{Z}$ are average log-spectrograms, the SSFs are obtained in the columns of $\mathbf{X}$. If the columns of the data matrix are the GSVs, their sketches are found in the columns of $\mathbf{X}$. The elements of $\mathbf{R}$, $R_{i,j}$, are taken as i.i.d. r.vs. sampled from a *p-stable distribution* [30].

A distribution $\mathcal{D}$ over $\mathbb{R}$ is called p-stable if there exists $p > 0$ such that for any n real numbers α_i, $i = 1, 2, \ldots, n$ and i.i.d. random numbers r_i drawn from $\mathcal{D}$, the r.v. $\sum_i \alpha_i\, r_i$ has the same distribution with the r.v. $(\sum_i |\alpha_i|^p)^{1/p}\, r$, where r is an r.v. having distribution $\mathcal{D}$. That is, if $R_{i,j}$ is sampled from a p-stable distribution, for any two average log-spectrograms (say the first two column of $\mathbf{Z}$) the differences $X_{i,1} - X_{i,2} = \sum_{j=1}^{d'} R_{i,j}(Z_{j,1} - Z_{j,2})$, $i = 1, 2, \ldots, d$, are also i.i.d. samples of a p-stable distribution. This implies that the projection can be used to recover an approximate value of the ℓ_p distance of the raw feature vectors computed in a space of reduced dimensions.

The most well-known stable distribution is the Gaussian distribution of zero mean and unit standard deviation $\mathcal{N}(0, 1)$, which is 2-stable. This distribution was used in [20]. However, the class of stable distributions is much wider, including heavy-tailed distributions as well [23]. For $p = 0.5$, one obtains the Levy distribution. The Cauchy distribution $f(r) = \frac{1}{\pi}\,\frac{1}{1+r^2}$ is 1-stable. Unfortunately, the aforementioned three distributions are the only cases for which closed form expressions of the probability density functions exist [31]. The probability density functions of symmetric p-stable distributions for different values of the tail constant p are plotted in Figure 5. However, samples can be drawn from any p-stable distribution [32]. Indeed, for $p \in (0, 2]$, $R_{i,j}$ are generated by [23]

$$R_{i,j} = \frac{\sin(p\theta)}{\cos^{1/p}\theta}\left(\frac{\cos(\theta(1-p))}{-\ln u}\right)^{\frac{1-p}{p}} \tag{4}$$

where θ is uniform on $[-\pi/2, \pi/2]$ and u is uniform on $[0, 1]$. (4) results for a zero skewness parameter

δ from the generic expression [31], [32]

$$R_{i,j} = \begin{cases} \gamma(p,\delta)\frac{\sin\, p\,(\theta+\theta_0)}{\cos^{\frac{1}{p}}\theta}\left(\frac{\cos(\theta-p(\theta+\theta_0))}{w}\right)^{\frac{1-p}{p}} & \text{if } p \neq 1 \\ \frac{2}{\pi}\left[(\frac{\pi}{2}+\delta\,\theta)\tan\theta - \delta\ln\left(\frac{\frac{\pi}{2}\,w\,\cos\theta}{\frac{\pi}{2}+\delta\,\theta}\right)\right] & \text{if } p = 1, \end{cases} \tag{5}$$

where $\gamma(p,\delta) = (1+(\delta\tan\frac{\pi\,p}{2})^2)^{\frac{1}{2p}}$, $\theta_0 = \frac{1}{p}\arctan(\delta\frac{\pi\,p}{2})$, and w is an exponentially distributed r.v. with mean 1. To generate samples from the Cauchy distribution (i.e., $p = 1$), both (4) and (5) yield $R_{i,j} = \tan\theta$.

Strictly speaking, the ℓ_p distance preserved by the p-stable random projections is a norm if $1 \leq p \leq 2$. In addition, the restricted isometry property (RIP) [33] for symmetric p-stable distributions has been proved only for $1 \leq p \leq 2$ [34]. The range of p will be confined to $1 \leq p \leq 2$, hereafter. Moreover, the projection matrix $\mathbf{R}$ is orthogonalized and the entries of $\mathbf{X} \in \mathbb{R}^{d\times n}$ 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 data matrix containing the average MFCC vectors is post-processed similarly.

III. Acquisition Device Identification via Sparse Representations

The problem of revealing the device identity of a test sketch given a number of labeled sketches from N acquisition devices is addressed based on the SRC [26]. Let $\mathbf{X}_i = [\mathbf{x}_{i,1}|\mathbf{x}_{i,2}|\ldots|\mathbf{x}_{i,n_i}] \in \mathbb{R}^{d\times n_i}$ denote the dictionary that contains n_i sketches, stemming from the ith device as column vectors (i.e., dictionary atoms). Given a test sketch $\mathbf{y} \in \mathbb{R}^d$ that comes from the ith device, one assumes that $\mathbf{y}$ is expressed as a linear combination of the atoms that are associated to the ith device, i.e.,

$$\mathbf{y} = \sum_{j=1}^{n_i} \mathbf{x}_{i,j}\, c_{i,j} = \mathbf{X}_i\, \mathbf{c}_i \tag{6}$$

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

If $\mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2|\ldots|\mathbf{X}_N] \in \mathbb{R}^{d\times n}$ is an overcomplete dictionary formed by concatenating n sketches, which stem from N acquisition devices[2], $\mathbf{y} \in \mathbb{R}^d$ in (6) is equivalently rewritten as $\mathbf{y} = \mathbf{X}\,\mathbf{c}$, where $\mathbf{c} = [\mathbf{0}^T|\ldots|\mathbf{0}^T|\mathbf{c}_i^T|\mathbf{0}^T|\ldots|\mathbf{0}^T]^T$ is the $n \times 1$ augmented coefficient vector, whose elements are zero except those associated with the ith device. Thus, the entries of $\mathbf{c}$ bear information about the device the test sketch $\mathbf{y} \in \mathbb{R}^d$ comes from.

[2]Obviously, $n = \sum_{i=1}^{N} n_i$.

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

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} \|\mathbf{c}\|_0 \quad \text{subject to } \mathbf{X}\mathbf{c} = \mathbf{y} \tag{7}$$

where $\|.\|_0$ is the ℓ_0 quasi-norm returning the number of the non-zero entries of a vector. Unfortunately, the solution of the problem (7) is NP-hard. An approximate solution to the problem (7) can be obtained by replacing the ℓ_0-norm with the ℓ_1-norm:

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} \|\mathbf{c}\|_1 \quad \text{subject to } \mathbf{X}\,\mathbf{c} = \mathbf{y} \tag{8}$$

where $\|.\|_1$ denotes the ℓ_1-norm of a vector. The ℓ_1 minimization (8) correctly recovers the ℓ_0 solution in (7) with $t+1$ nonzero elements, if $\psi\, d < t < \lfloor \frac{d+1}{3} \rfloor$ for some small constant ψ [35]. That is, as long as the number of nonzero elements of the ℓ_0 solution is a small fraction of the dimension d, the solution of (7) is equivalent to the solution of (8), which can be obtained by standard linear programming methods in polynomial time. This is a necessary and sufficient condition related to the neighborliness of the convex polytope spanned by the columns of $\mathbf{X}$[3] [35], while the RIP gives sufficient conditions only [26]. It is worth noting that by projecting the data using a p-stable matrix, the dictionary $\mathbf{X}$ obeys the RIP for $1 \leq p \leq 2$ [34].

A test sketch is classified as follows. The coefficient vector $\mathbf{c}^*$ is obtained by solving (8). Ideally, $\mathbf{c}^*$ contains non-zero entries in positions associated with the dictionary atoms (i.e., columns of $\mathbf{X}$) stemming from a single device, so that one can easily assign the test sketch $\mathbf{y}$ 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 sketch is classified to the device class that minimizes the residual $\|\mathbf{y} - \mathbf{X}\,\delta_i(\mathbf{c})\|_2$, where $\delta_i(\mathbf{c}) \in \mathbb{R}^n$ is a vector, whose nonzero entries are associated to the ith device only [26].

[3]That is, when $\mathbf{X}$ maps all t dimensional facets of the unit ℓ_1 ball in $\mathbb{R}^n$ to the facets of the convex polytope spanned by the columns of $\mathbf{X}$, which belongs to $\mathbb{R}^d$.

A. Robustness to sparse modeling errors

A more general model for a test sketch $\mathbf{y} \in \mathbb{R}^d$ allows for an error vector $\mathbf{e} \in \mathbb{R}^d$, having a fraction of its elements nonzero [26], i.e.,

$$\mathbf{y} = \mathbf{X}\,\mathbf{c} + \mathbf{e} = [\mathbf{X}|\mathbf{I}] \begin{bmatrix} \mathbf{c} \\ \mathbf{e} \end{bmatrix}, \tag{9}$$

where $\mathbf{I} \in \mathbb{R}^{d \times d}$ is the identity matrix. The error vector $\mathbf{e}$ elements have arbitrary magnitude and cannot be ignored or treated by replacing the equality constraint with an ℓ_2-norm regularization term $\|\mathbf{y} - \mathbf{X}\mathbf{c}\|_2 \leq \varepsilon$ in (8). Let $\hat{\mathbf{X}} = [\mathbf{X}|\mathbf{I}] \in \mathbb{R}^{d \times (n+d)}$ be the extended dictionary and $\mathbf{v} = [\mathbf{c}^T\ \mathbf{e}^T]^T$.

Next, the sparsest solution $\mathbf{v}^*$ is found by solving the extended ℓ_1 minimization problem:

$$\mathbf{v}^* = \arg\min_{\mathbf{v}} \|\mathbf{v}\|_1 \quad \text{subject to } \hat{\mathbf{X}}\,\mathbf{v} = \mathbf{y}. \tag{10}$$

Once the sparsest solution $\mathbf{v}^* = [\hat{\mathbf{c}}^T\ \hat{\mathbf{e}}^T]^T$ is determined, each test sketch is classified to the device class that minimizes the modified residual $\|\mathbf{y} - \hat{\mathbf{e}} - \mathbf{X}\,\delta_i(\hat{\mathbf{c}})\|_2$.

IV. Experimental Evaluation

Experiments were conducted on two databases, namely, a subset of the LLHDB [19] as in [2] for landline telephone handset identification and the MOBIPHONE for source cell-phone identification.

The subset of the LLHDB consists of speech recordings from 53 speakers (24 male speakers and 29 female speakers) acquired by 8 landline telephone handsets. 4 of telephone handsets are carbon-button (CB1-CB4) and the remaining 4 are electrect (EL1-EL4). Each speaker utters 10 sentences from the TIMIT database [29]. The recording are approximately 3 s long. 2 of the 10 sentences (i.e., SA1 and SA2) are read by every speaker and the remaining 8 sentences are different. The latter were used for device identification.

The MOBIPHONE contains 21 cell-phones of various models from 7 different brands. The brands and models of the cell-phones are listed in Table I. For 12 male speakers and another 12 female speakers, randomly chosen from the TIMIT database, 10 utterances were recorded by the various cell-phones. These 10 utterances per speaker were concatenated in a single 30 s long recording.

Two sets of experiments were conducted for landline telephone handset identification. The first set of experiments follows the 2-fold cross-validation set-up, while the second one partitions the recordings so that the speakers, whose utterances were included in the test set, were left out from the training set. A

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

Class Name	Brand and Model	Class Name	Brand and Model
HTC1	HTC desire c	APL1	iPhone5
HTC2	HTC sensation	S1	Samsung E2121B
LG1	LG GS290	S2	Samsung E2600
LG2	LG L3	S3	Samsung GT-18190 mini
LG3	LG Optimus L5	S4	Samsung GT-N7100 (galaxy note2)
LG4	LG Optimus L9	S5	Samsung Galaxy GT19100 s2
N1	Nokia 5530	S6	Samsung Galaxy Nexus S
N2	Nokia C5	S7	Samsung e1230
N3	Nokia N70	S8	Samsung s5830i
SE1	Sony Ericson c902	V1	Vodafone joy 845
SE2	Sony Ericson e510i		

third set of experiments was conducted for cell-phone identification using disjoint partitions of training and test speakers.

A. Landline telephone handset identification using 2-fold cross-validation

The first set of experiments follows the 2-fold cross-validation set-up for device identification in [2], [3]. The 3392 recordings were divided into two disjoint sets that are balanced in the number of recordings, the number of speakers, and their gender). The first subset of 1696 recordings was used to derive the training dictionaries, while the second subset of 1696 recordings was exploited for testing.

The top identification accuracies are summarized in Table II, when the SSFs or the MFCCs are classified by the SRC solving (8) for Cauchy and Gaussian random projections, the linear SVM [27], [28], and the NN with the cosine similarity measure. By inspecting Table II, it is seen that the SSFs identify the acquisition device committing less errors than the MFFCs, no matter which classifier is employed. Moreover, the SSFs achieve state-of-the-art identification accuracy, if they are fed to either the SVM or the SRC classifier for both stable distributions considered. The latter classifier achieves the highest identification accuracy (i.e., 94.99%), when Gaussian random projections are used. The SRC outperforms also the SVM, when the MFCCs are employed.

The performance of the SRC solving (8) and the SVM in telephone handset identification as a function of feature dimension d for SSFs obtained by several values of p is depicted in Fig. 6. The best accuracy

TABLE II
TOP TELEPHONE HANDSET IDENTIFICATION ACCURACIES ACHIEVED BY THE SSFS FOR CAUCHY AND GAUSSIAN RANDOM PROJECTIONS AND THE MFCCS, WHEN THE SRC SOLVING (8), THE LINEAR SVM, AND THE NN ARE EMPLOYED.

Features	**Feature dimension**	**Classifier**	**Accuracy (%)**
SSFs (Cauchy)	800	SRC	**94.72**
SSFs (Cauchy)	800	SVM	94.66
SSFs (Cauchy)	775	NN	83.78
SSFs (Gaussian)	700	SRC	**94.99**
SSFs (Gaussian)	800	SVM	94.66
SSFs (Gaussian)	850	NN	85.08
MFCCs	23	SRC	89.79
MFCCs	23	SVM	87.35
MFCCs	23	NN	81.95
MFCCs- based Gaussian supervector [2]	N/A	SVM	93.20

(i.e., 95.08%) was obtained for the SRC with $p = 1.5$ and $d = 850$. Clearly, for $d > 175$ the SRC outperforms the best result reported in [2], demonstrating the robustness of the proposed approach.

The performance of the SRC solving (10) is compared to that of the SVM in telephone handset identification for SSFs of several dimensions d obtained using several values of p in Fig. 7. The best accuracy (i.e., 95.46%) was obtained for the SRC for Cauchy random projections (i.e., $p = 1$) and $d = 550$. The SVM attained the top accuracy of 94.96% for the SSFs of the same dimension $d = 550$ and a stable projection with $p = 1.3$. It is seen that the best accuracies were obtained for random projections employing p-stable distributions different than the Gaussian (i.e., $p \neq 2$).

In order to check if the accuracy differences are statistically significant, we apply the approximate analysis in [36]. Assume that the accuracies ϖ_1 and ϖ_2 are binomially distributed random variables. If $\hat{\varpi}_1, \hat{\varpi}_2$ denote the empirical accuracies, and $\overline{\varpi} = \frac{\hat{\varpi}_1 + \hat{\varpi}_2}{2}$, the hypothesis $H_0 : \varpi_1 = \varpi_2 = \overline{\varpi}$ is tested at 95% level of significance. The accuracy difference has variance $\beta = 2\frac{\overline{\varpi}(1-\overline{\varpi})}{M}$, where M is the number of test recordings (i.e., 1696). For $\varphi = 1.65\sqrt{\beta}$, if $\hat{\varpi}_1 - \hat{\varpi}_2 \geq \varphi$, we reject H_0 with risk 5% of being wrong. The aforementioned analysis yields that the performance gain between the SRC or the SVM employing the SSFs and that reported in [2] is statistically significant ($\varphi = 1.35\%$), while the accuracy differences between the SRC and the SVM are not.

B. Landline telephone handset identification leaving test speakers out from training

Two sketches of features were tested for landline telephone handset identification, namely the SSFs and the sketches of the GSVs. The GSVs were extracted resorting to a GMM-UBM for 128 and 256 components trained on the training subset of the TIMIT. The 3392 recordings of the LLHDB were split into two disjoint sets as follows. The utterances of 15 randomly chosen female speakers and 12 randomly chosen male speakers were exploited in order to derive the dictionaries of the sketches, while the utterances of the remaining 14 female speakers and the 12 male speakers were used to extract the test sketches. By doing so, the training and test sets include 1728 and 1664 recordings, respectively. Three such random permutations of 15 among the 29 female speakers and 12 among 24 male speakers where made and the resulting accuracies were averaged. To the best of the author's knowledge, such an evaluation is performed for first time in the literature.

The linear SVM was found to yield statistically significant accuracy gains in telephone handset identification over the SRC solving either (8) or (10). Due to space limitations, comparisons are made between the linear SVM and the SRC solving (8). Fig. 8 depicts the accuracy of each classifier as a function of feature dimension d of the SSFs obtained for various p. The best accuracy (i.e., 81.95%) was obtained for the SVM with Gaussian random projections (i.e., $p = 2$). For $d = 625$ and $p = 1.6$, the SRC attained an average accuracy of 78.53%. The accuracy difference 3.42% exceeds the threshold $\varphi = 2.28\%$ of the approximate analysis, which implies a statistically significant performance gain.

Table III summarizes the accuracies measured for the three classifiers, when the sketches of the GSVs were employed for 128 and 256 components in the GMMs. The threshold on the difference of the SRC and SVM accuracies that guarantees statistically significant performance gains is also included in Table III. It is seen that for 128 components the performance gain of the SVM over the SRC is not statistically significant. On the contrary, the performance gain of the SVM over the SRC is statistically significant for 256 components. Both the SRC and the SVM offer statistically significant performance gains over the NN classifier. The best accuracy in Table III is statistically significant over that reported in [2].

C. Source cell-phone identification leaving test speakers out from training

Various sketches of features were tested for cell-phone identification, namely the sketches of the GSVs, the SSFs, and the average MFCCs. The GSVs were extracted by exploiting a GMM-UBM with 64 and 128 components trained on the entire training subset of the TIMIT as well as the dialect region DR2 of the training subset of the TIMIT. In the MOBIPHONE, two disjoint subsets of 252 recordings were

TABLE III
BEST TELEPHONE HANDSET IDENTIFICATION ACCURACIES ACHIEVED BY THE SKETCHES OF GSVS WHEN THE SRC SOLVING (10), THE LINEAR SVM, AND THE NN ARE EMPLOYED.

K	**Classifier**	p	d	**Accuracy (%)**	$\varphi(\mathrm{SRC}, \mathrm{SVM})$
128	SRC	1.2	775	93.63	1.36
	SVM	1.5	775	94.11	
	NN	1.2	400	74.34	
256	SRC	1.3	400	89.66	1.52
	SVM	1.6	775	**94.65**	
	NN	1.8	850	74.70	

created, that are balanced in the number of recordings as well as speakers and gender in order to MAP update the GMM-UBM mean vectors. The sketches of GSVs had size $d \leq 120$. In addition, GSVs were extracted by training a GMM with a number of components K ranging from 1 to 5 on the MFCCs extracted from the recordings of the MOBIPHONE. In the latter case, the GSVs were made by stacking the component mean vectors and the vector comprising the elements of diagonal component covariance matrices on their main diagonal. For such GSVs, various sizes $d \leq d'/2 = (K \times 22)$ were tested. The SSFs of size $d \leq 850$ as well as the average MFFCs were also employed for cell-phone identification. The random projections reduced the size of the average MFFCs to $d \leq 11$.

The best identification accuracies are summarized in Table IV, when the SRC solving (10), the linear SVM, and the NN with the cosine similarity measure classified the sketches of features. By inspecting Table IV, it is seen that a perfect classification was achieved by the SRC and the SVM applied to GSVs, exploiting the GMM-UBM. When the entire training set of TIMIT was used in the GMM-UBM training, the NN achieves a perfect classification as well. An almost perfect classification was achieved by the NN when the DR2 dialect region was employed to train the GMM-UBM. It is worth mentioning that the top performance was consistently obtained for $p \leq 1.3$ and $d \leq 100$. For the GSVs training on the MOBIPHONE data with 1 or 2 components, a high accuracy was measured for all classifiers. The marginal accuracy differences are not statistically significant at 95% level of significance. The SRC outperformed the SVM, when the SSFs were employed. In this case, the accuracy difference between any pair of classifiers is statistically significant at 95% level of significance. For the sketches of the average MFFCs, the SVM achieved the top accuracy among the three classifiers. Although not directly comparable, the accuracies obtained here are higher than those reported in [3].

TABLE IV
BEST SOURCE CELL-PHONE IDENTIFICATION ACCURACIES ACHIEVED BY THE SRC SOLVING (10), THE LINEAR SVM, AND THE NN FOR VARIOUS SKETCHES OF FEATURES.

Sketches of Features	**Classifier**	**Accuracy (%)**
GSVs using a GMM-UBM trained on the full training subset of TIMIT ($K = 64, 128$)	SRC	**100**
	SVM	**100**
	NN	**100**
GSVs using a GMM-UBM trained on the training subset DR2 of TIMIT ($K = 64$)	SRC	**100**
	SVM	**100**
	NN	99.60
GSVs using a GMM-UBM trained on the training subset DR2 of TIMIT ($K = 128$)	SRC	**100**
	SVM	**100**
	NN	98.41
GSV with 1 GMM component trained on the MOBIPHONE	SRC	99.21
	SVM	98.41
	NN	98.02
GSV with 2 GMM components trained on the MOBIPHONE	SRC	96.83
	SVM	95.24
	NN	94.05
SSFs	SRC	98.81
	SVM	96.03
	NN	94.84
average MFFCs	SRC	96.82
	SVM	97.22
	NN	96.03

V. CONCLUSIONS

The sketches of features have been demonstrated to capture the intrinsic trace of the acquisition device in vectors of small size, speeding up classification and resulting to memory savings. By employing a proper p stable distribution to randomly project raw feature vectors of large size to sketches, very high rates have been obtained for landline telephone handset identification in the LLHDB as well as source cell-phone identification in the MOBIPHONE by either a sparse representation-based classifier or a support vector machine. The concept of correntropy [37], as a generalized similarity measure between two arbitrary random variables, can be exploited in the re-formulation of the identification problem.

Acknowledgments. This work has been supported by the Cost Action IC 1106 “Integrating Biometrics and Forensics for the Digital Age”. The author would like to thank Mr. Stamatios Samaras for the collection of the MOBIPHONE database during his B.Sc. Thesis.

REFERENCES

[1] H. Farid, “Digital image forensics,” *Scientific American*, vol. 6, no. 298, pp. 66–71, 2008.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[22] C. Kotropoulos, "Telephone handset identification using sparse representations of spectral feature sketches," in *Proc. 1st Int. Workshop Biometrics and Forensics*, Lisbon, Portugal, 2013.

[23] P. Indyk, "Stable distributions, pseudorandom generators, embeddings, and data stream computation," *Journal of the ACM*, vol. 53, no. 3, pp. 307–323, 2006.

[24] W. M Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

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

[26] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

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

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

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

[30] V. Zolotarev, *One Dimensional Stable Distributions*, vol. 65, Translations of Mathematical Monographs, American Mathematical Society, Providence, RI, USA, 1986.

[31] G. R. Arce, *Nonlinear Signal Processing*, J. Wiley & Sons, Hoboken, NJ, USA, 2005.

[32] J. P. Nolan, *Stable Distributions*, Birkhauser, 2002.

[33] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[34] D. Otero and G. R. Arce, "Generalized restricted isometry property for alpha-stable random projections," in *Proc. 2011 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Prague, The Czech Republic, 2011, pp. 3676–3679.

[35] D. Donoho, "For most large underdetermined systems of equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 7, pp. 907–934, 2006.

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

[37] R. He, W.-S. Zheng, T. Tan, and Z. Sun, "Half-quadratic based iterative minimization for robust sparse representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2013, in press.

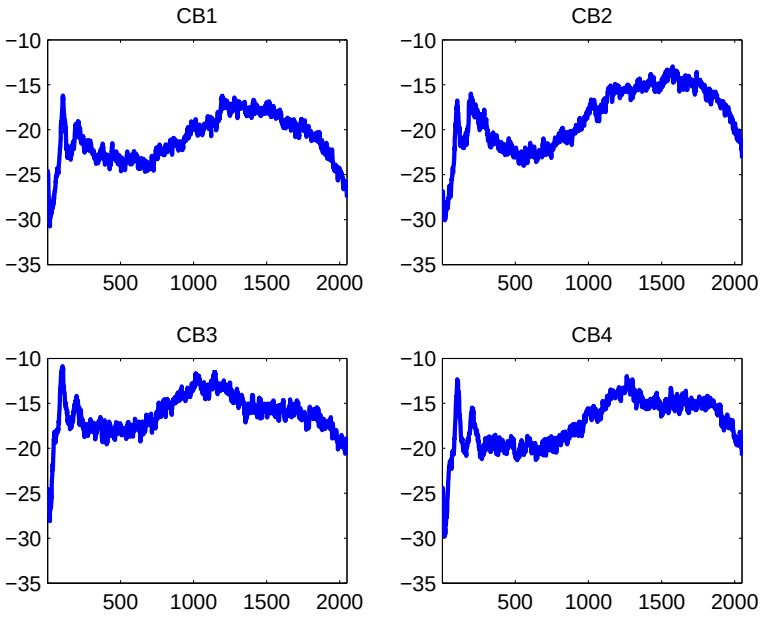


Fig. 1. Average log-spectrograms for the 4 carbon button landline telephone handsets in the LLHDB (Horizontal axis: frequency index; Vertical axis: average log-spectrogram value).

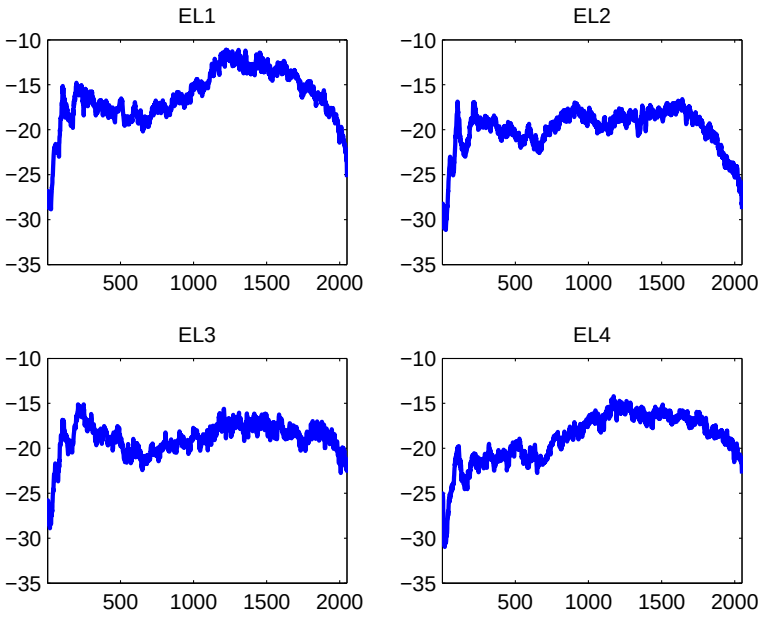


Fig. 2. Average log-spectrograms for the 4 electret landline telephone handsets in the LLHDB (Horizontal axis: frequency index; Vertical axis: average log-spectrogram value).

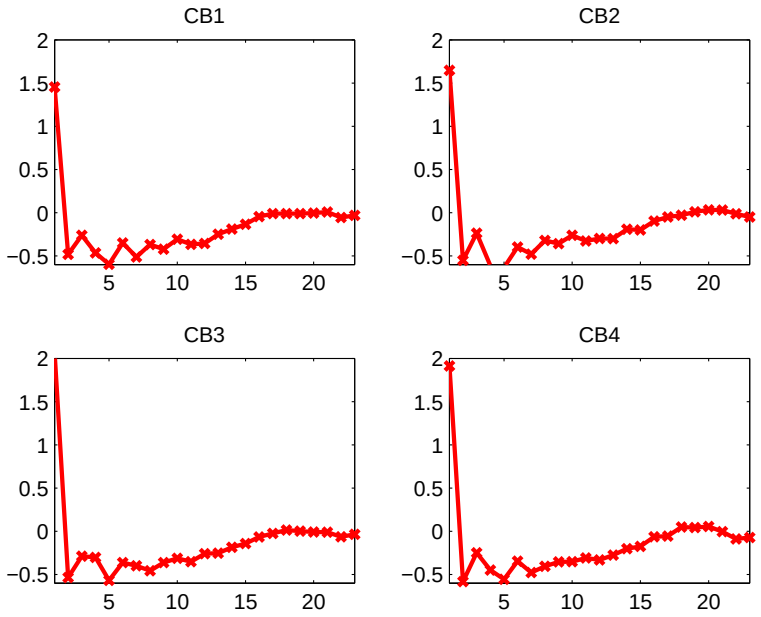


Fig. 3. Average MFCCs for the 4 carbon button landline telephone handsets in the LLHDB (Horizontal axis: MFCC index; Vertical axis: average MFCC value).

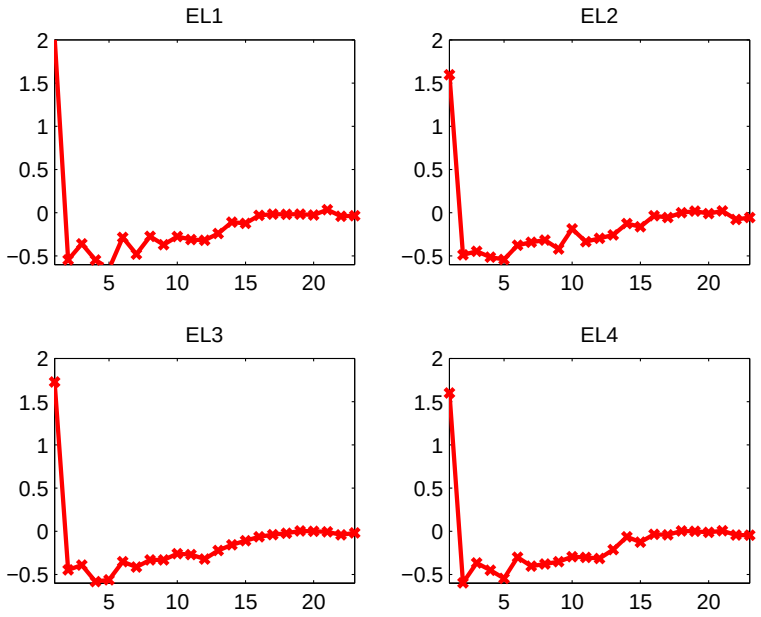


Fig. 4. Average MFCCs for the 4 electret landline telephone handsets in the LLHDB (Horizontal axis: MFCC index; Vertical axis: average MFCC value).

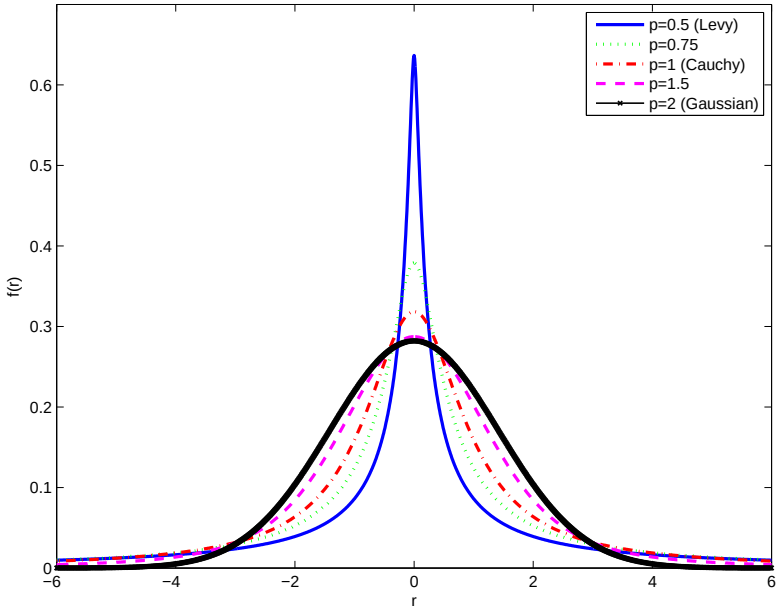
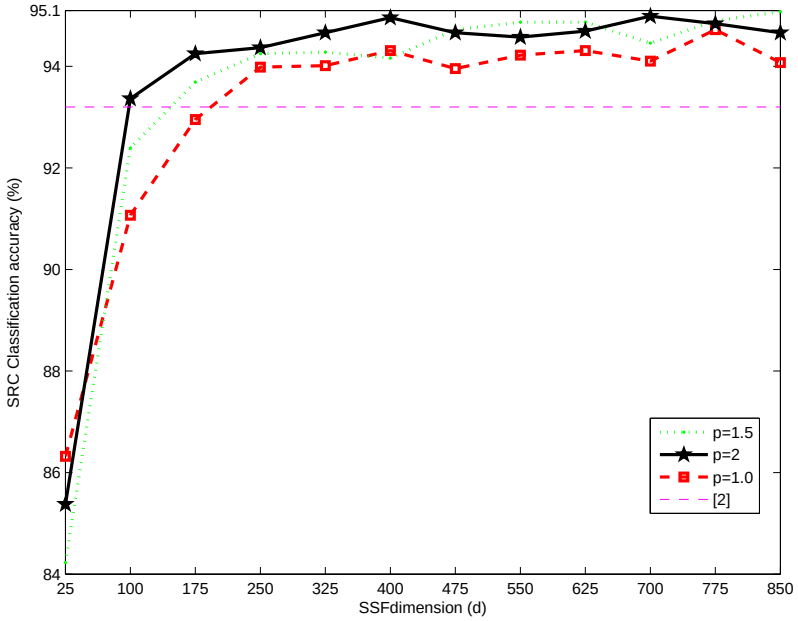
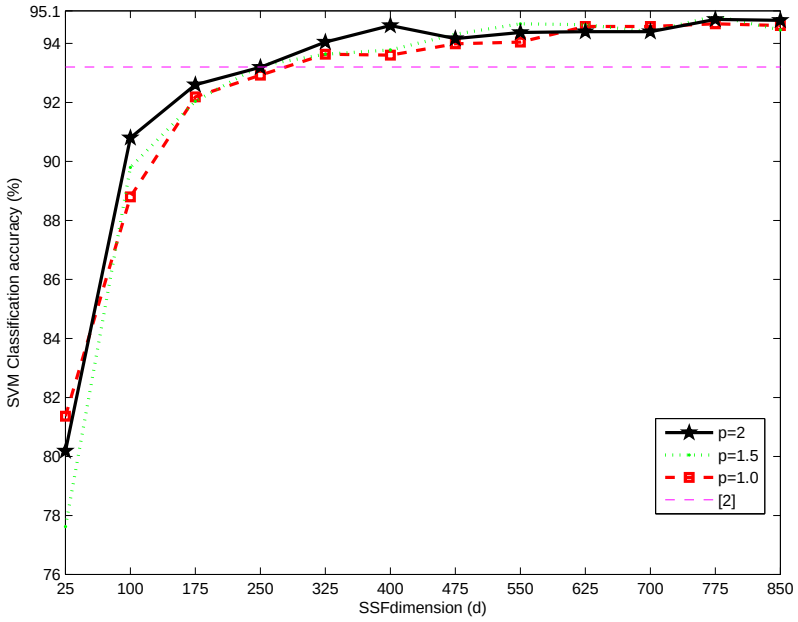


Fig. 5. Probability density functions of symmetric p-stable distributions for different values of the tail constant p.



(a)



(b)

Fig. 6. Telephone handset identification accuracy versus the SSF dimension d achieved by (a) the SRC solving (8) and (b) the SVM for various p.

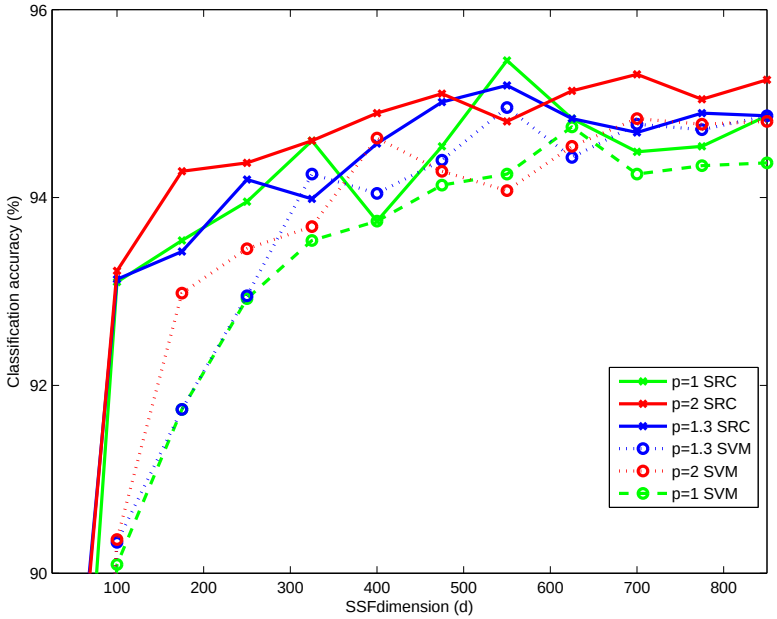


Fig. 7. Telephone handset identification accuracy versus the SSF dimension d achieved by the SRC solving (10) and the SVM for various p.

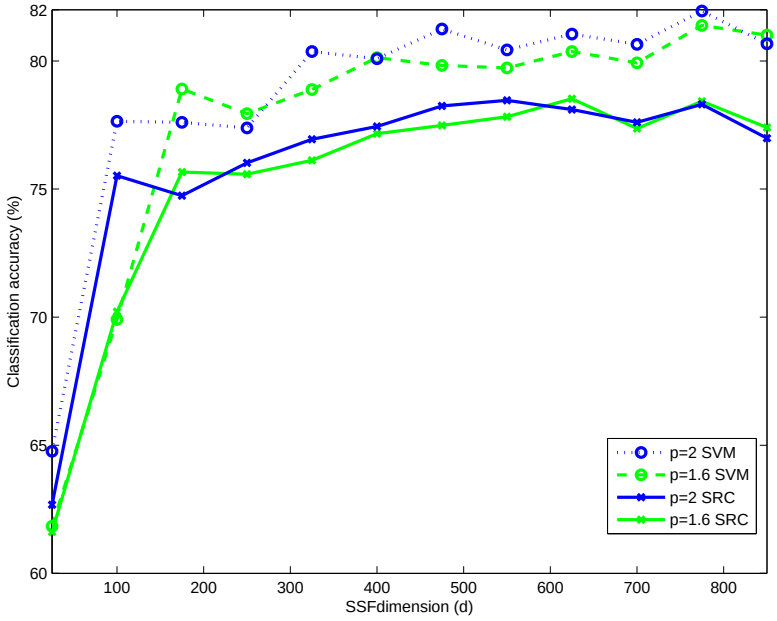


Fig. 8. Telephone handset identification accuracy versus the SSF dimension d achieved by the SRC solving (8) and the SVM for various p.