TELEPHONE HANDSET IDENTIFICATION USING SPARSE REPRESENTATIONS OF SPECTRAL FEATURE SKETCHES

Constantine Kotropoulos

Department of Informatics
Aristotle University of Thessaloniki
Box 451, Thessaloniki 54124, GREECE
costas@aiia.csd.auth.gr

ABSTRACT

Speech signals convey useful information for the recording devices used to capture them. Here, acquisition device identification is studied using the *sketches of spectral features* (SSFs) as intrinsic fingerprints. The SSFs are extracted from the speech signal by first averaging its spectrogram along the time axis and then by mapping the resulting mean spectrogram into a low-dimension space, such that the "distance properties" of the high-dimensional mean spectrograms are preserved. Such a mapping results by taking the inner product of the mean spectrogram with a vector of independent identically distributed random variables drawn from a p-stable distribution. By applying a sparse-representation based classifier to the SSFs, state-of-the-art identification accuracy exceeding 95% has been measured on a set of 8 telephone handsets from Lincoln-Labs Handset Database (LLHDB).

Index Terms— Digital speech forensics, sketches, spectral features, sparse representation.

1. 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 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]. Codec identification has attracted the interest of the forensics community. Studies 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]. The authentication of speakers' environment has been investigated [9, 10, 11, 12]. The effectiveness of Hidden Markov Model-based phone recognition for forensic voice comparison has been evaluated in terms of both validity (accuracy) and reliability (precision) in [13]. A few automatic acquisition device identification systems have been developed. For instance, a method for the classification of 4 microphones has been proposed in [10] that was further improved thanks to a proper fusion strategy [11]. 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 the Naive Bayes classifier at a short-time frame level. Accuracies in the order of 60-75% have been reported. Rank level fusion was shown to increase classification accuracy to 100% [11]. The identification of 8 landline telephone handsets and 8 microphones is addressed in [2]. In particular, the intrinsic characteristics of the device are captured by a template constructed by concatenating the mean vectors 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 (SVM) classifier was employed. Recently, a robust system for the identification of cell-phones has been proposed in [3]. 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, the blind-passive method for landline telephone handset identification introduced in [14] is elaborated further. This method resorts on suitable feature extraction from speech recordings and their sparse representation, enabling to trace the recording device. Here, the *sketches of spectral features* (SSFs) are proposed as intrinsic fingerprints suitable for device identification. The SSFs are extracted from the speech signal first by averaging its spectrogram along the time axis and second by mapping the mean spectrogram into a low-dimension space, such that the "distance properties" of the high-dimensional mean spectrograms are preserved. Such a mapping can be obtained by taking the inner product of the mean spectrogram with a vector of independent identically distributed (i.i.d.) random variables drawn from a p-stable distribution [15]. A special case of the SSFs are the random spectral features used in [14]. The SSFs form an overcomplete dictionary of basis signals for devices' intrinsic traces. This dictionary is exploited then for *sparse representation-based classification* (SRC) [16]. If sufficient training speech recordings are available for each device, it is possible to express the SSFs extracted from a recording captured by an unknown

978-1-4673-4989-5/13/$31.00

(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 vector of test SSFs the device identity (ID) the dictionary atoms weighted by non-zero coefficients are associated with.

The proposed method is tested for the identification of 8 telephone handsets by conducting experiments on the Lincoln-Labs Handset Database (LLHDB) [17], 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 [18] 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 fingerprints, no matter which classifier is employed. Meanwhile, the proposed device identification method yields an accuracy of 95.02%, outperforming the state-of-the-art method [2] on the LLHDB dataset.

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

2. ACQUISITION DEVICE FINGERPRINTS

The majority of features employed in 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, 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.

Let us first extract the spectrogram 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 2048 samples. Next, the logarithm of the spectrogram is calculated and is averaged along the time axis, yielding a 2048-dimensional mean spectrogram.

Denote the data matrix by $\mathbf{Z} \in \mathbb{R}^{2048\times n}$ containing the mean spectrograms of n recordings. The dimensionality of the mean spectrograms is reduced to $d < 2048$ by pre-multiplying $\mathbf{Z}$ with a projection matrix $\mathbf{R} \in \mathbb{R}^{d\times 2048}$ yielding $\mathbf{X} = \mathbf{R}\,\mathbf{Z}$. The elements of $\mathbf{R}$, $R_{i,j}$, can be taken as i.i.d. random variables sampled from a *p-stable distribution* [19]. A distribution $\mathcal{D}$ over $\mathbb{R}$ is called p-stable if there exists $p \geq 0$ such that for any n real numbers α_i, $i = 1, 2, \ldots, n$ and i.i.d. random variables r_i drawn from $\mathcal{D}$, the random variable $\sum_i \alpha_i\, r_i$ has the same distribution as the variable $(\sum_i |\alpha_i|^p)^{1/p}\, r$, where r is a random variable having distribution $\mathcal{D}$. That is, if we sample $R_{i,j}$ from a p-stable distribution, for any two mean spectrograms (say the first two column of $\mathbf{Z}$) the differences $X_{i,1} - X_{i,2} = \sum_{j=1}^{2048} R_{i,j}(Z_{j,1} - Z_{j,2})$, $i = 1, 2, \ldots, d$, 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 norm of the original spectrograms 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 [14]. However, the class of stable distributions is much wider, including heavy-tailed distributions as well [15]. For example, the Cauchy distribution $f(r) = \frac{1}{\pi}\,\frac{1}{1+r^2}$ is 1-stable. In general for $p \in (0, 2]$, $R_{i,j}$ can be generated by [20]

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

where θ is uniform on $[-\pi/2, \pi/2]$ and u is uniform on $[0, 1]$. 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 columns of $\mathbf{X}$ are the SSFs that are used for acquisition device identification.

The MFCCs are considered as baseline features [2]. They encode the frequency content of the speech signal by parameterizing the rough shape of 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 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 SSFs.

3. ACQUISITION DEVICE IDENTIFICATION VIA SPARSE REPRESENTATION

The problem of revealing the device identity of a vector of SSFs given a number of labeled SSFs from N acquisition devices is addressed based on the SRC [16].

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 SSFs stemming from the ith device as column vectors (i.e., dictionary atoms). Given a vector of test SSFs $\mathbf{y} \in \mathbb{R}^d$ 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.,

$$\mathbf{y} = \sum_{j=1}^{n_i} \mathbf{a}_{i,j}\, c_{i,j} = \mathbf{A}_i\, \mathbf{c}_i \tag{2}$$

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

Next, let $\mathbf{A} = [\mathbf{A}_1|\mathbf{A}_2|\ldots|\mathbf{A}_N] \in \mathbb{R}^{d\times n}$ be an overcomplete dictionary formed by concatenating n SSFs, which stem from N acquisition devices[1] Thus, $\mathbf{y} \in \mathbb{R}^d$ in (2) is equivalently rewritten as $\mathbf{y} = \mathbf{A}\,\mathbf{c}$, 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 vector of SSFs $\mathbf{y} \in \mathbb{R}^d$ comes from.

Since the device ID of a test vector of SSFs 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 SSFs $\mathbf{y} \in \mathbb{R}^d$, 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.,

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

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

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

(3) is NP-hard. An approximate solution to the problem (3) can be obtained by replacing the ℓ_0 norm with the ℓ_1 norm:

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

where $\|.\|_1$ denotes the ℓ_1 norm of a vector. In [21], 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 SSFs 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 SSFs $\mathbf{y}$ 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 SSF is classified to the device class that minimizes the residual $\|\mathbf{y} - \mathbf{A}\,\delta_i(\mathbf{c})\|_2$, where $\delta_i(\mathbf{c}) \in \mathbb{R}^n$ is a new vector, whose nonzero entries are associated to the ith device only [16].

4. EXPERIMENTAL EVALUATION

Experiments were conducted on the same subset of the Lincoln-Labs Handset Database (LLHDB) [17] as in [2]. 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 [2], a stratified 2-fold cross-validation is employed.

Table 1. Best telephone handset identification accuracies achieved by the SSFs and the MFCCs, when the SRC, 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

The best identification accuracies are summarized in Table 1, when the SSFs or the MFCCs are classified by the SRC [16], the linear SVM [18], and the NN with the cosine similarity measure. By inspecting Table 1, it is clear that the SSFs are able to 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%) on the LLHDB, when Gaussian random projections are used. The SRC outperforms also the SVM, when the MFCCs are employed.

The performance of the SRC and the SVM in telephone handset identification on the LLHDB as a function of feature dimension (i.e., d) for SSFs obtained by several values of p is depicted in Fig. 1. The best accuracy (i.e., 95.08%) is obtained for the SRC with $p = 1.5$ and $d = 850$. Clearly, for $d > 175$ and $p \geq 1$ the SRC outperforms the best result reported in [2], demonstrating the robustness of the proposed approach.

In order to check if the accuracy differences are statistically significant, we apply the approximate analysis in [22]. Let us 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 $\zeta = 1.65\sqrt{\beta}$, if $\hat{\varpi}_1 - \hat{\varpi}_2 \geq \zeta$, 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 ($\zeta = 1.35\%$), while the accuracy differences between the SRC and the SVM are not.

It is worth noting that by projecting the data onto an orthogonal p-stable matrix, the dictionary $\mathbf{A}$ obeys the restricted isometry property (RIP) for a certain, appropriate order (say S) [23]. When this property holds, $\mathbf{A}$ approximately preserves the Euclidean length of S-sparse SSFs, 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.

5. CONCLUSIONS

The SSFs have been 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 SSFs over the MFCCs for device characterization, yielding a state-of-the-art performance in recognizing 8 telephone handsets from the LLHDB.

Acknowledgments. This work has been supported by the Cost Action IC 1106 "Integrating Biometrics and Forensics for the Digital Age".

6. REFERENCES

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

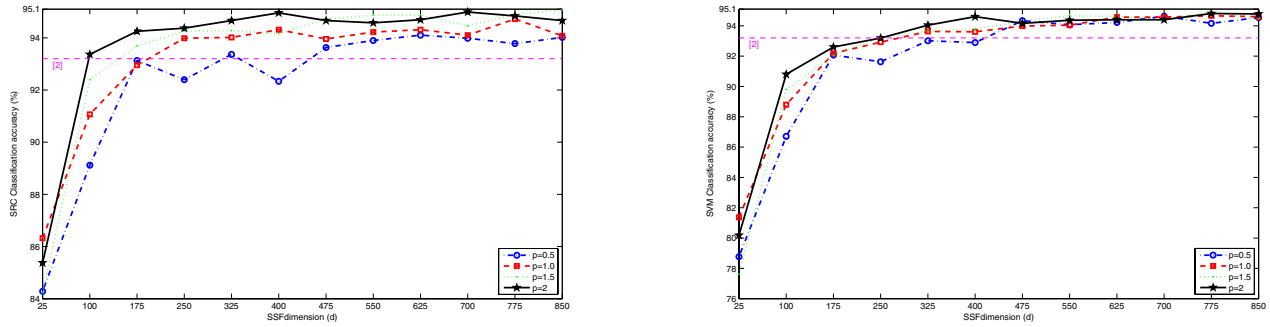
[2] D. Garcia-Romero and C. Y. Espy-Wilson, "Automatic acquisition device identification from speech recordings," in *Proc. 2010 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Dallas, TX, USA, 2010, pp. 1806–1809.

[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.*



(a) (b)

Fig. 1. Telephone handset identification accuracy versus the SSF dimension d achieved by (a) the SRC and (b) the SVM on the LLHDB for various p.

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] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

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

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

[21] 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.

[22] 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.

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