SOURCE CELL PHONE MATCHING FROM SPEECH RECORDINGS BY SPARSE REPRESENTATION AND KISS METRIC

Ling Zou, Qianhua He, Jichen Yang and Yanxiong Li

School of Electronic and Information Engineering
South China University of Technology, Guangzhou 510640
zou.ling@mail.scut.edu.cn

ABSTRACT

Source recording device matching from two speech recordings is a new and important problem of digital media forensics. It aims to answer the question that whether or not two speech recordings are recorded by the same recording device. In this study we propose a source cell phone matching scheme. The Gaussian supervector (GSV) based on Mel-frequency cepstral coefficients (MFCCs) is extracted from the speech recording and is sparse represented with respect to a dictionary learned by K-SVD algorithm. The reduced-dimensional sparse representation coefficient is utilized to characterize the intrinsic fingerprint of the recording device. Then, KISS metric learning based similarity matching is conducted on a pair of fingerprints extracted from the two speech recordings. Evaluation experiments were conducted on a database of speech recordings recorded by 14 cell phones. The experimental results demonstrated the feasibility of the proposed scheme.

Index Terms— Digital audio forensics, Source recording device matching, KISS metric learning, Sparse representation.

1. INTRODUCTION

Recognition of the source recording device from the speech recordings would prove useful in the court for establishing the authenticity of speech recordings presented as evidence [1, 2]. Over the past several years, source recording device recognition has gained more attention. The literature is largely clustered into a few specific problems such as source microphone identification [3-10], source telephone handset identification [10-16], source mobile device identification [16-21] and source cell phone verification [21, 22]. For example, Hanilçi *et al.* [18] firstly tried to identify the brands and models of cell phones from the speech recording using Mel-frequency cepstral coefficients (MFCCs) and support vector machine (SVM). More recently, Zou *et al.* presented a source cell phone verification scheme based on sparse representation [22].

Most studies focus on the source recording device recognition problem. To our best knowledge, there are no studies which have focused on the source recording device matching problem. As illustrated in Fig. 1, given two speech recordings, source recording device matching aims to determine whether or not the two speech recordings are acquired by the same recording device without having to know the specific device unit. This problem is full of realistic significance in the forensic context. For instance, in some cases the recording device is unavailable or unusable (e.g., has been damaged) for the court, also in some cases we don't care about the specific recording device and only want to know whether the two speech recordings are recorded by the same device. This problem is similar to the source recording device verification problem in [22]. However, the difference is that, in the latter case, the claimed recording device is supposed to be known and available which means that we can obtain sufficient examples acquired by the recording device. It will benefit for the subsequent verification task whereas this doesn't hold for the source recording device matching task because there are only two speech recordings available here. Therefore, the available information quantity for the matching task is less and the methods for source recording device verification may not be well suited for source recording device matching problem.

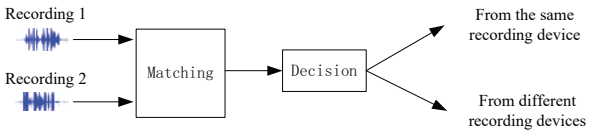


Fig. 1. Illustration of source recording device matching problem.

Motivated by the forensic significance and characteristic of source recording device matching, on the other hand, considering the fact that the wide availability of cell phone signifies that there will be lots of evidences in the form of digital speech recordings taken to court. Thus, in this study we take cell phone as the representative recording device and try to address this new problem, i.e., source cell phone matching from speech recordings. Given two speech recordings, for each speech recording, the Gaussian supervector (GSV) is extracted. It is sparse represented with respect to a dictionary learned by some kind of learning algorithm (here K-SVD). The reduced-dimensional sparse representation coefficients are utilized to represent the intrinsic fingerprint of the recording device. Then, the similarity matching is conducted by computing the KISS metric [23-25] based distance between the fingerprints extracted from the two speech recordings. The performance of the proposed scheme is evaluated on a database of speech recordings recorded by 14 cell phones.

The rest of the paper is organized as follows. In section 2, the method of this study is described. Section 3 details the experimental set up in this paper. Section 4 presents the experimental results and discussion. Finally, future work and conclusions are summarized in Section 5.

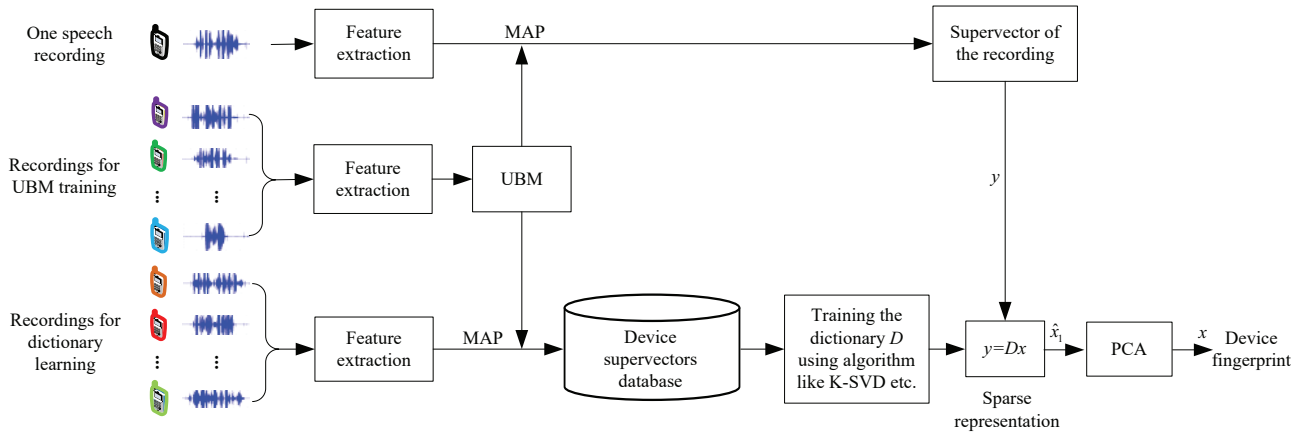


Fig. 2. Block diagram of extraction of recording device fingerprint from a speech recording.

2. METHODS

2.1. Gaussian supervector

GSV has shown success in representation of the intrinsic fingerprint of a recording device from speech recordings [10]. Extracting GSV from a speech recording is conducted as follows: Suppose that $\lambda_{\text{UBM}} = \{\omega_i, \mu_i, \Sigma_i\}_{i=1}^{K}$ is a diagonal covariance universal background model (UBM) with K mixtures, given an speech recording and suppose that the corresponding feature vectors (MFCCs in this study) extracted from it is $X = \{x_t\}_{t=1}^{T}$, then, the means adapted Gaussian mixture model (GMM) is updated from the UBM by *maximum a posteriori* (MAP) [26]. Suppose that $\lambda_a = \{\omega_i, \mu_i^a, \Sigma_i\}_{i=1}^{K}$ and $\lambda_b = \{\omega_i, \mu_i^b, \Sigma_i\}_{i=1}^{K}$ are the obtained means adapted GMMs corresponding to two speech recordings, then the Kullback-Leibler (KL) divergence kernel is defined as the corresponding inner product of the two GSVs which is a concatenation of the weighted mean vectors from each mixture of the GMM [27]:

$$K(\lambda_a, \lambda_b) = \sum_{i=1}^{K} (\sqrt{w_i}\Sigma_i^{-(1/2)}\mu_i^a)^T (\sqrt{w_i}\Sigma_i^{-(1/2)}\mu_i^b). \quad (1)$$

2.2. Sparse representation based device fingerprint extraction

In the statistical signal processing field, given an overcomplete dictionary $D \in \mathbb{R}^{M\times N}$ ($M<<N$) which is made up of N base elements (a.k.a. atoms), an input signal $y \in \mathbb{R}^M$ can be represented by the sparse linear combination of these atoms: $y = Dx$. Here, $x \in \mathbb{R}^N$ is the sparse representation coefficient with majority of the entries are zero.

Given a training vectors set $Y = \{y_i\}_{i=1}^{N}$ (here $y_i \in \mathbb{R}^M$ is the GSV in this study), the K-SVD algorithm [28] searches for the best possible dictionary D for the sparse representation of Y by solving

$$\min_{D,X} \left\{ \|Y - DX\|_2^2 \right\} \quad \text{subject to} \quad \forall i \quad \|x_i\|_0 \le T_0 \quad (2)$$

where X is the corresponding sparse representations to Y and T_0 is the sparsity constraint. Once D is determined, the input vector y can be sparsely represented with respect to D using the basis pursuit (BP) approach [29] by solving

$$\hat{x}_1 = \arg\min \|x\|_1 \quad \text{subject to} \quad \|y - Dx\|_2 \le \varepsilon \quad (3)$$

where $\hat{x}_1$ is the obtained sparse representation coefficient.

The dimensionality of the sparse representation coefficient is usually high. To make it tractable for distance metric learning algorithm, we use PCA [30] to conduct dimensionality reduction. Suppose that $U_{\text{PCA}} \in \mathbb{R}^{N\times d}$ ($d < N$) is the obtained feature extractor after principle component analysis to original high dimensional feature space, then the low dimensional feature can be obtained by

$$x = U_{\text{PCA}}^T \hat{x}_1 \quad (4)$$

where $x \in \mathbb{R}^d$ is the low-dimensional representation of high-dimensional sparse coefficient $\hat{x}_1$. x is used as the intrinsic recording device fingerprint in this study. The whole extraction procedure for this type of device fingerprint is illustrated in Fig. 2.

2.3. KISS metric based device fingerprint matching

Suppose that x_a and x_b are a pair of feature samples, Mahalanobis distance metric measures the distance between the pair of samples as:

$$D_M^2(x_a, x_b) = (x_a - x_b)^T M (x_a - x_b) \quad (5)$$

where M is a positive semidefinite matrix. From the viewpoint of KISS metric learning [23], the similarity of a pair of samples (x_a, x_b) can be determined using the log likelihood ratio based on a statistical inference perspective:

$$\delta(x_a, x_b) = \log\left(\frac{p(x_a, x_b \mid H_0)}{p(x_a, x_b \mid H_1)} \right) \quad (6)$$

where H_0 and H_1 stand for the hypothesis that the pair of samples are dissimilar and similar respectively. Assuming $x_{ab} = x_a - x_b$ has zero-mean Gaussian distribution, then the problem can be casted into

$$\delta(x_{ab}) = \log\left(\frac{p(x_{ab} \mid H_0)}{p(x_{ab} \mid H_1)} \right) = \log\left(\frac{f(x_{ab} \mid \theta_0)}{f(x_{ab} \mid \theta_1)} \right) = \log\left(\frac{\frac{1}{\sqrt{2\pi\left|\Sigma_{y_{ab}=0}\right|}} \exp(-1/2 x_{ab}^T \Sigma_{y_{ab}=0}^{-1} x_{ab})}{\frac{1}{\sqrt{2\pi\left|\Sigma_{y_{ab}=1}\right|}} \exp(-1/2 x_{ab}^T \Sigma_{y_{ab}=1}^{-1} x_{ab})} \right) \quad (7)$$

where f is the pdf with parameters θ_1 for hypothesis H_1 and θ_0 for hypothesis H_0. Under the Gaussian assumption, the parameters of the two distributions are $\theta_1 = (0, \Sigma_{y_{ab}=1})$ and $\theta_0 = (0, \Sigma_{y_{ab}=0})$ where

$$\Sigma_{y_{ab}=1} = \sum_{y_{ab}=1} (x_a - x_b)(x_a - x_b)^T, \tag{8}$$

$$\Sigma_{y_{ab}=0} = \sum_{y_{ab}=0} (x_a - x_b)(x_a - x_b)^T. \tag{9}$$

After taking the log, equation (7) can be re-formulated as

$$\begin{aligned}\delta(x_{ab}) = {} & x_{ab}^T \sum\nolimits_{y_{ab}=1}^{-1} x_{ab} + \log\left(\left|\Sigma_{y_{ab}=1}\right|\right) \\ & - x_{ab}^T \sum\nolimits_{y_{ab}=0}^{-1} x_{ab} - \log\left(\left|\Sigma_{y_{ab}=0}\right|\right).\end{aligned} \tag{10}$$

Neglecting the constant terms, we get

$$\begin{aligned}\delta(x_{ab}) &= x_{ab}^T (\Sigma_{y_{ab}=1}^{-1} - \Sigma_{y_{ab}=0}^{-1}) x_{ab} \\ &= (x_a - x_b)^T M_{\mathrm{KISS}} (x_a - x_b).\end{aligned} \tag{11}$$

In this study, suppose that x_1 and x_2 are the recording device fingerprints extracted from two speech recordings as shown in Fig. 2. Once the matrix M_{KISS} of the KISS metric learning is obtained on a training set, we can calculate the KISS metric based distance between x_1 and x_2 as

$$D_M^2(x_1, x_2) = (x_1 - x_2)^T M_{\mathrm{KISS}} (x_1 - x_2) \underset{<}{\geq} \theta. \tag{12}$$

Besides the KISS metric, the conventional Mahalanobis distance (M is set to the covariance matrix), the cosine kernel metric:

$$\frac{\langle x_1, x_2 \rangle}{\|x_1\| \|x_2\|} \underset{<}{\geq} \theta \tag{13}$$

and the correlation metric:

$$\frac{\langle (x_1 - \bar{x}_1), (x_2 - \bar{x}_2) \rangle}{\|x_1 - \bar{x}_1\| \|x_2 - \bar{x}_2\|} \underset{<}{\geq} \theta \tag{14}$$

are also considered for reference.

3. EXPERIMENTAL SETUP

We evaluated the proposed source cell phone matching scheme on a database of speech recordings from 14 cell phones [18, 22]. Table 1 shows the detail of the cell phones in this database. The database was collected by two means and each resulted in a subset. The first subset was obtained by playing a subset of the TIMIT corpus through all the 14 cell phones in a silent environment using a loudspeaker. This subset contains 24 speakers and there are 10 sentences for each speaker. Thus there are 240 speech recordings for each cell phone (120 recordings for training UBM and dictionary, 60 recordings for KISS metric learning and the remaining 60 recordings for final matching testing). The duration for each recording is approximately 3 seconds. Apart from TIMIT subset, the second subset was collected by recording a speaker speaking into the above 14 cell phones a passage of approximately 10 minutes in the same room. Then, each recording was evenly segmented into 200 recordings each with the duration of approximately 3 seconds. Thus there are 200 speech recordings for each cell phone (100 recordings for training UBM and dictionary, 50 recordings for KISS metric learning and the remaining 50 recordings for final matching testing). This subset is referred to as LIVE hereafter.

When we carried out experiments on TIMIT subset, the training part of LIVE subset (collecting the training part of all cell phones) were utilized for training the UBM and vice versa. For the current experimental subset, the training part of this subset was utilized for learning the dictionary. Specifically, the statistics of the final matching number of the two subsets in one experiment are listed in Table 2.

Table 1. Brands and models of the cell phones in the experimental database.

BRAND	MODEL
SAMSUNG	E250, E250, D900
NOKIA	2730, 6500, 3600, 3600, 6670
MOTOROLA	Q
SONY	W880, W880, K750I
LG	KE970
HP	IPAQ514

Table 2. Statistics of matched and unmatched trials in the matching experiment.

Subsets	Device number	Recordings per class	Matched testing	Unmatched testing
TIMIT	14	60	24780	327600
LIVE	14	50	17150	227500

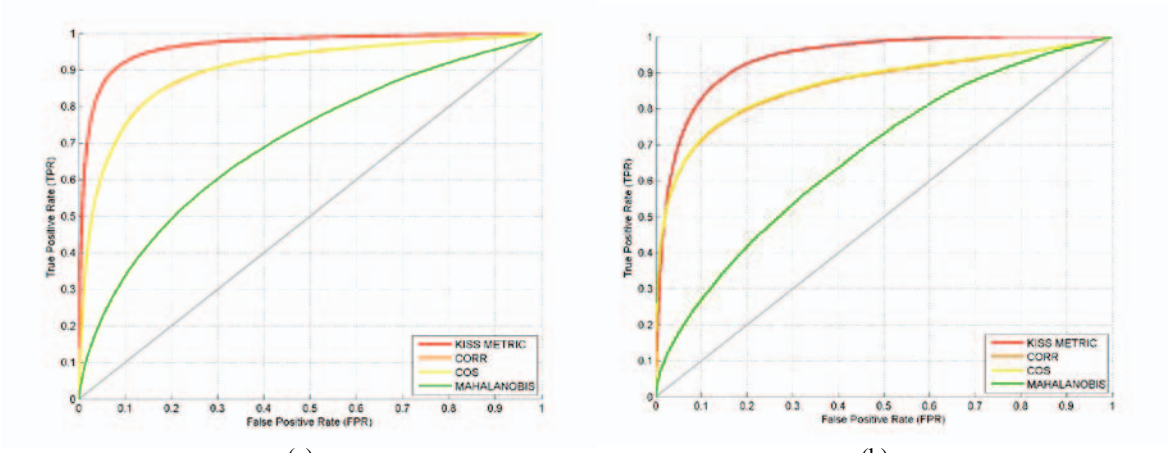
Table 3. EERs (in %) for different matching methods (here, the fingerprint dimensionality is 100).

Matching metrics	TIMIT	LIVE
KISS metric	**8.83**	**13.09**
correlation	16.59	20.13
cosine	16.58	19.86
conventional Mahalanobis distance	35.13	38.16

For each speech recording, the whole utterance, including speech segments and non-speech segments, was segmented into frames by a 30 ms Hamming window with an overlap of 50%. 12 MFCCs were computed using 27 triangular filters with c0 excluded. The MFCCs were concatenated with the energy feature and resulted in a 13-dimensional feature vector. The number of mixture components in UBM was set to 32, therefore the dimensionality of the GSV was of 416. The size of the learned dictionary is set to 416×1260 in this study, this set also guaranteed that the dictionary is redundant and overcomplete.

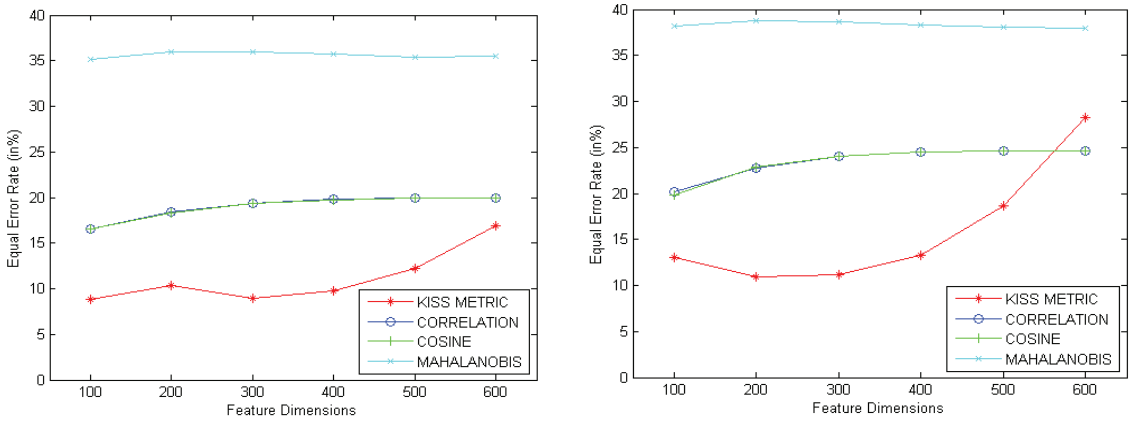
4. RESULTS AND DISCUSSION

Firstly, we carry out matching experiments on two subset using KISS metric and other three types of metrics. The matching number is in accordance with Table 2. The dimensionality of the fingerprint is set to 100. The experimental results are illustrated in Table 3 and Fig. 3. Table 3 shows the achieved Equal Error Rates (EERs) when utilizing different matching methods. It can be observed that the KISS metric obtains the best matching performance compared to the other three methods especially on the TIMIT subset where the EERs relatively decreased about 46.7% compared to correlation and cosine metric, and 74.8% compared to conventional Mahalanobis distance. It shows the effectiveness of the KISS metric based matching to this problem. The corresponding Receiver Operator Characteristic (ROC) curve is illustrated in Fig. 3. In addition, we can also find that the cosine kernel metric and the correlation metric yield so close results that



(a) (b)

Fig. 3. The matching results when the fingerprint dimensionality is set to 100. (a) ROC curves obtained on TIMIT subset and (b) ROC curves obtained on LIVE subset.



(a) (b)

Fig. 4. The EERs for different fingerprint dimensionality. (a) results on TIMIT subset (b) results on LIVE subset.

the ROC curves for them are almost overlapped.

For evaluating the influence of reduced dimensionality to the matching performance, we also conducted experiments when different subspace dimensionality of device fingerprint are utilized. It can be observed from Fig. 4 that the KISS metric is much influenced by the feature dimensionality compared to cosine metric, correlation metric and the conventional Mahalanobis distance metric. The overall tendency of the matching results when KISS metric is utilized is that the larger the feature dimensionality, the more arising the matching error. The reason may be due to that a higher dimensional feature subspace means that more data will be needed for learning an accurate KISS metric matrix M_{KISS}. However, the amount of training data is limited and fixed in this study, therefore, the M_{KISS} will tend to be more accurately learned in a relatively low-dimensional feature subspace. Nevertheless, it can also be observed from Fig. 4 that the KISS metric still outperforms the other three metrics almost on all feature dimensionalities in the experiments.

This is a preliminary study of source cell phone matching problem. It should be noted that this study focus only on one type of recording device, i.e., cell phone. However, it is possible that the proposed scheme could be extended to other types of recording device like tablet, voice recorder etc.

5. CONCLUSIONS

In this study we try to address a new problem of source cell phone matching from speech recordings and propose a scheme which is based on sparse representation and KISS metric learning. The KISS metric outperforms the cosine kernel metric, the correlation metric and the conventional Mahalanobis distance to this problem. We also investigate the influence of fingerprint dimensionality to the matching performance and find that the KISS metric is much influenced compared to other three types of metrics when the size of the training set for KISS metric learning is fixed. The larger feature dimensionality requires more training data for learning an accurate KISS metric matrix *M*. To sum up, we propose a source cell phone matching scheme in this study. Future work includes extending the experimental database and further enhancing the matching performance.

6. ACKNOWLEDGEMENT

This work was supported by the National Nature Science Foundation of China (Grant No. 61571192, 61301300); Natural Science Foundation of Guangdong Province (Grant No. 2015A030313600) and the Fundamental Research Funds for the Central Universities, SCUT (Grant No. 2015ZM143).

7. REFERENCES

[1] H. Malik and H. Zhao, "Recording environment identification using acoustic reverberation," in *Proc. ICASSP*, 2012, pp. 1833–1836.

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

[3] C. Kraetzer, A. Oermann, J. Dittmann, and A. Lang, "Digital Audio Forensics: A First Practical Evaluation on Microphone and Environment Classification," in *Proc. 9th Workshop on Multimedia and Security*, Dallas, TX, USA, 2007, pp. 63–74.

[4] 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*, 2009, pp. 49–56.

[5] C. Kraetzer, K. Qian, M. Schott, and J. Dittmann, "A context model for microphone forensics and its application in evaluations," *Media Watermarking, Security, and Forensics III*, vol. 7880, 2011.

[6] R. Buchholz, C. Kraetzer, and J. Dittmann, "Microphone classification using Fourier coefficients," in *Lecture Notes in Comput. Sci*. Berlin/Heidelberg, Germany: Springer, 2010, vol. 5806/2009, pp. 235–246.

[7] H. Malik and J. Miller, "Microphone identification using higher-order statistics," in *Proc. AES 46th Conf. Audio Forensics 2012*, Denver, CO, USA, 2012.

[8] O. Eskidere, "Source microphone identification from speech recordings based on a Gaussian mixture model," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 22, no. 3, pp. 754–767, 2014.

[9] D. Garcia-Romero and C. Espy-Wilson, "Speech forensics: Automatic acquisition device identification," *J. Acoust. Soc. Am.,* vol. 127, no. 3, pp. 2044–2044, 2010.

[10] D. Garcia-Romero and C. Espy-Wilson, "Automatic acquisition device identification from speech recordings," in *Proc. ICASSP*, 2010, pp. 1806–1809.

[11] D. Garcia-Romero and C. Espy-Wilson, "Automatic acquisition device identification from speech recordings," *J. Audio Eng. Soc.*, vol. 124, no. 4, pp. 2530–2530, 2009.

[12] D. A. Reynolds, "HTIMIT and LLHDB: speech corpora for the study of handset transducer effects," in *Proc. ICASSP*, Munich, Germany, 1997, vol. 2, pp. 1535–1538.

[13] 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. 91–96.

[14] 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.

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

[16] C. Kotropoulos, "Source phone identification using sketches of features," *Biometrics, IET*, vol. 3, no. 2, pp. 75–83, 2014.

[17] C. Kotropoulos and S. Samaras, "Mobile phone identification using recorded speech signals," in *Digital Signal Processing (DSP), 2014 19th International Conference on*, 2014, pp. 586–591.

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

[19] L. Zou, J. C. Yang, and T. S. Huang, "Automatic cell phone recognition from speech recordings," in *Signal and Information Processing (ChinaSIP), 2014 IEEE China Summit & International Conference on*, 2014, pp. 621–625.

[20] M. Jahanirad, A. W. A. Wahab, N. B. Anuar, M. Y. I. Idris, and M. N. Ayub, "Blind source mobile device identification based on recorded call," *Engineering Applications of Artificial Intelligence*, vol. 36, pp. 320–331, 2014.

[21] C. Hanilçi and T. Kinnunen, "Source cell-phone recognition from recorded speech using non-speech segments," *Digital Signal Processing*, vol. 35, pp. 75–85, 2014.

[22] L. Zou, Q. H. He, and X. H. Feng, "Cell phone verification from speech recordings using sparse representation," in *Proc. ICASSP*, 2015, pp. 1787–1791.

[23] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf.Comput. Vision Pattern Recogn.*, Jun. 2012, pp. 2288–2295.

[24] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li, "Person re-identification by regularized smoothing KISS metric learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 10, pp. 1675–1685, Oct. 2013.

[25] D. Tao, L. Jin, Y. Wang, and X. Li, "Person reidentification by minimum classification error-based KISS metric learning," *IEEE Trans. Cybern.*, vol. 45, no. 2, pp. 242–252, Feb. 2015.

[26] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

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

[28] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.

[29] A. Yang, Z. Zhou, A. Ganesh, S. Sastry, and Y. Ma, "Fast L1-minimization algorithms for robust face recognition," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3234–3246, 2013.

[30] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, Aug. 2006.