SPARSITY BASED ROBUST SPEAKER IDENTIFICATION USING A DISCRIMINATIVE DICTIONARY LEARNING APPROACH

Christos Tzagkarakis and Athanasios Mouchtaris

Department of Computer Science, University of Crete and
Institute of Computer Science, Foundation for Research and Technology - Hellas
FORTH-ICS
Heraklion, Crete, Greece
{tzagarak,mouchtar}@ics.forth.gr

ABSTRACT

Speaker identification is a key component in many practical applications and the need of finding algorithms, which are robust under adverse noisy conditions, is extremely important. In this paper, the problem of text-independent speaker identification is studied in light of classification based on sparsity representation combined with a discriminative dictionary learning technique. Experimental evaluations on a small dataset reveal that the proposed method achieves a superior performance under short training sessions restrictions. In specific, the proposed method achieved high robustness for all the noisy conditions that were examined, when compared with a GMM universal background model (UBM-GMM) and sparse representation classification (SRC) approaches.

Index Terms— speaker identification, sparse representation, discriminative dictionary learning, K-SVD

1. INTRODUCTION

Speaker recognition constitutes an essential part of distinct emerging applications. Ranging from the control of financial transactions, the entrance into safe or reserved areas and buildings, and the information retrieval from speech databases [1], to the most modern technologies of speaker-tracking during a teleconference, speaker diarization for meetings, and speech-aided systems in ambient intelligence environments [2, 3]. This necessitates strongly the design and development of efficient speaker recognition algorithms characterized by increased robustness in diverse environments and conditions.

Speaker recognition can be categorized into speaker verification and speaker identification. In speaker verification, a speaker claims to be of a certain identity and his/her voice is used to verify this claim. On the other hand, speaker identification is the task of determining an unknown speaker's identity. Speaker recognition methods can also be divided into text-dependent and text-independent methods. The former requires the speaker to provide utterances of keywords or sentences, the same text being used for both training and recognition. In text-independent recognition, the decision does not rely on a specific text being spoken.

In this paper, we focus on text-independent speaker identification. A commonly used approach, in order to estimate the identity of a speaker, is to assign each candidate speaker from a given database to a specific speaker model describing consistently the extracted speech features. During the identification process, the system decides for the speaker's identity based on the closest matching of the test utterance against all speaker models. State-of-the-art methods follow a universal speaker modeling approach, such as, Gaussian mixture models (GMMs) [4, 5] using all the available training data to build a model before the acquisition of the test sample.

In the current work, we study the problem of noise robust text-independent speaker identification under the assumption of short training sessions. This is crucial in applications where the voice data are obtained in highly noisy environments and it is often not feasible to have large amounts of training data from all the speakers. Towards this direction, a discriminative learning approach is introduced. The problem is faced under a joint learning perspective, where an overcomplete dictionary is learned, resulting in highly discriminative sparse codes, along with a linear classifier. A speech corpus of twelve speakers is used for the identification evaluation towards the direction of examining applications consisting of a moderate number of speakers (*e.g.,* teleconference meetings).

The rest of the paper is organized as follows: Section 2 overviews the most recent methods on sparsity based classification methods for speech signals, while Section 3 describes in brief the sparse representation classification (SRC) approach. In Section 4 we analyze the proposed approach for noise robust speaker identification using a discriminative dictionary learning algorithm. The experimental evaluation and comparison of the performance of the proposed technique with state-of-the-art methods is described in Section 5. Finally, Section 6 summarizes the main conclusions and gives directions for future work.

2. PRIOR WORK ON SPARSITY BASED CLASSIFICATION FOR SPEECH SIGNALS

The concept of sparse representation (or sparse coding) comes as an alternative solution to the universal data models, which do not generalize well for limited training data. Prior work on classification of speech signals have recently provided promising results. The main focus is given on representing an input test sample as a sparse linear combination of an overcomplete matrix, the so-called *dictionary*, whose columns consist of a set of basis functions, usually referred to as atoms. In [6], robust speech recognition is achieved by modeling noisy speech signals as a sparse linear combination of speech and noise *exemplars* (spectro-temporal representations spanning multiple time-frames of the speech signal). A similar approach is fol-

This research has been co-financed by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thalis-DISFER

lowed in [7], where a combination of large vocabulary continuous speech recognition techniques with small vocabulary tasks results in low phonetic error rates.

Sparse codes may also serve as a new type of feature vectors to be given as input in a typical classifier. More specifically, a gradient descent-based dictionary learning approach is adopted in [8] to learn the redundant matrix related with the training data. This comes in combination with a multilayer perceptron classifier, which is applied on the generated sparse codes for phoneme recognition. The same task is also studied in [9]. An orthogonal matching pursuit-based (OMP) dictionary learning technique is applied and the obtained sparse codes are further used for classification by means of a support vector machine (SVM) classifier. A phone recognition approach employing hidden Markov models (HMM) is examined in [10], using sparse codes which take advantage of the phonetic labels information as additional features during the recognition process. Moreover, the sparse codes feature extraction is followed by sparse discriminant analysis to perform speaker recognition in [11], while in [12] SRC is used for the same task using GMM mean supervectors as feature vectors on clean speech data taken from TIMIT speech corpus.

Dictionary learning techniques can be applied for learning the best dictionary that gives the most discriminative sparse codes for classification. The work in [13] showed that a satisfactory speaker verification performance can be achieved by applying a supervised K-SVD algorithm for learning an appropriate discriminative dictionary. Motivated by the successful application of K-SVD for face and object categorization [14], our proposed method addresses the problem of text-independent speaker identification by extending our previous work [15]. Here, we adopt a discriminative dictionary learning approach, which is applied on noise robust speaker identification under the assumption of short training speech utterances.

Our proposed method learns an overcomplete dictionary, resulting in highly discriminative sparse codes, along with a linear classifier. This estimation is performed in a joint fashion by imposing additional constraints on the associated objective function in order to produce similar sparse codes for those training samples belonging to the same speaker. This is in contrast to recently introduced sparsity-based methods [6, 7, 8, 9, 10, 11, 12], which do not treat jointly the estimation of the dictionary, the sparse codes, and the classifier parameters. On the other hand, in [13], a method was suggested to learn jointly only the dictionary and the sparse codes. To the best of our knowledge, this is the first study on noise robust speaker identification, which tackles the problem from such a threefold joint learning perspective.

3. SPARSE REPRESENTATION CLASSIFICATION FOR SPEAKER IDENTIFICATION

In this section, the sparse representation classification (SRC) method is briefly described, which was also employed in our previous work [15] for noise robust speaker identification using short training and testing speech data. Let S be the total number of speakers in our database. Then, for each speaker, a matrix $\mathbf{V}_i$ is constructed based on the feature vectors extracted from the i-th speaker as follows:

$$\mathbf{V}_i = [\mathbf{v}_{i,1}|\mathbf{v}_{i,2}|\cdots|\mathbf{v}_{i,n_i}] \in \mathbb{R}^{d\times n_i},\ i = 1,\ldots,S, \tag{1}$$

where the column vector $\mathbf{v}_{i,j}$ denotes the j-th d-dimensional feature vector of the i-th speaker, and n_i is the number of training feature vectors for the i-th speaker. The total number of training feature vectors in our database equals $N_{tr} = n_1 + \ldots + n_S$. In the present work, mel-frequency coefficients have been used as feature vectors (ref. Section 5).

In a speaker identification application, the goal is to infer correctly the identity of an unknown speaker, given a new test sample (feature vector) $\mathbf{x}_t \in \mathbb{R}^{d\times 1}$. In the following, let $\mathbf{x}_t$ be a feature vector, which is extracted from the i-th speaker. Then, it can be expressed as a linear combination of the training samples associated with this speaker as follows:

$$\mathbf{x}_t = c_{i,1}\mathbf{v}_{i,1} + c_{i,2}\mathbf{v}_{i,2} + \cdots + c_{i,n_i}\mathbf{v}_{i,n_i} = \mathbf{V}_i\,\mathbf{c}_i, \tag{2}$$

where $\mathbf{c}_i = \{c_{i,j}\}_{j=1}^{n_i}$ is the vector of coefficients of the representation of $\mathbf{x}_t$ in terms of the columns of $\mathbf{V}_i$.

The overall training data matrix $\mathbf{V}$ is formed by concatenating all the training data matrices $\mathbf{V}_i$, $i = 1,\ldots,S$,

$$\begin{aligned}\mathbf{V} &= [\mathbf{v}_{1,1}|\cdots|\mathbf{v}_{1,n_1}|\mathbf{v}_{2,1}|\cdots|\mathbf{v}_{2,n_2}|\cdots|\mathbf{v}_{S,1}|\cdots|\mathbf{v}_{S,n_S}]\\ &= [\mathbf{V}_1|\mathbf{V}_2|\cdots|\mathbf{V}_S] \in \mathbb{R}^{d\times N_{tr}}\ .\end{aligned} \tag{3}$$

By combining (2) and (3), $\mathbf{x}_t$ can be expressed in terms of the overall training data matrix $\mathbf{V}$, namely, $\mathbf{x}_t = \mathbf{V}\mathbf{c}$, where

$$\mathbf{c} = [0,\ \ldots,\ 0,\ c_{i,1},\ c_{i,2},\ \ldots,\ c_{i,n_i},\ 0,\ \ldots,\ 0] \in \mathbb{R}^{N_{tr}\times 1} \tag{4}$$

denotes the coefficients vector, hereafter called the *sparse code*, whose elements are all zero except for those associated with the i-th speaker. Notice that, the larger the number of speakers S is, the sparser the sparse code $\mathbf{c}$ will be.

Given the training data matrix $\mathbf{V}$ and the new feature vector (test sample) $\mathbf{x}_t$, the following optimization problem can be solved through the orthogonal matching pursuit (OMP) [16] algorithm in order to obtain an estimate of $\mathbf{c}$,

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \|\mathbf{x}_t - \mathbf{V}\mathbf{c}\|_2\ ,\ \text{s.t.}\ \|\mathbf{c}\|_0 = K, \tag{5}$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm, $\|\cdot\|_0$ is the ℓ_0 (pseudo)norm, which is defined as the number of non-zero elements of a given vector and K denotes the number of iterations of the algorithm or, equivalently, the number of non-zero elements in $\hat{\mathbf{c}}$.

In the ideal case, the indices of the non-zero entries of $\hat{\mathbf{c}}$ will correspond to those columns of $\mathbf{V}$ associated with the i-th speaker, and thus, the test sample $\mathbf{x}_t$ will be assigned correctly to that speaker. However, due to potential modeling errors and/or noise-corrupted data, in practice there may be also several non-zero entries of small amplitude in $\hat{\mathbf{c}}$, which correspond to multiple speakers. To overcome this drawback, we define for each speaker i an indicator function $\delta_i : \mathbb{R}^{N_{tr}} \to \mathbb{R}^{N_{tr}}$ such that the only non-zero entries of vector $\delta_i(\hat{\mathbf{c}}) \in \mathbb{R}^{N_{tr}}$ are from the i-th speaker, and this procedure is repeated S times for each speaker. As a result, for a given speaker i we can approximate $\hat{\mathbf{x}}_t^i = \mathbf{V}\delta_i(\hat{\mathbf{c}})$ and assign the test sample to the speaker with the minimum residual between $\mathbf{x}_t$ and $\hat{\mathbf{x}}_t^i$ as

$$i^* = \arg\min_i \|\mathbf{x}_t - \mathbf{V}\delta_i(\hat{\mathbf{c}})\|_2\ ,\ i = 1,\ldots,S. \tag{6}$$

This process is performed for each frame of the speech signal of the speaker to be identified, and the final class, that is, the speaker's identity, is estimated by means of a majority voting approach applied on a predefined set of frames. In other words, the unknown speaker is assigned the class to which most of the frames of his/her speech signal are classified in using (6).

4. DISCRIMINATIVE DICTIONARY SPARSE CODING BASED ON K-SVD

In this section, the method of discriminative dictionary sparse coding based on a (class) label-consistent K-SVD is analyzed, which constitutes the key component of our proposed approach. This method, which was introduced in the framework of face and object recognition [14] and to our knowledge is now applied for a first time in the field of speaker identification. We apply the method in the context of *noisy* conditions using *small training* data sessions.

Following the notation of Section 3, the sparse coding optimization problem expressed by (5) can be extended to the following *dictionary learning* optimization problem:

$$\hat{\mathbf{D}}, \hat{\mathbf{C}} = \arg\min_{\mathbf{D},\mathbf{C}} \|\mathbf{V} - \mathbf{D}\mathbf{C}\|_F^2 , \quad \text{s.t. } \|\mathbf{c}_j\|_0 = K , \ \forall j = 1, \ldots, N_{tr} , \tag{7}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, $\mathbf{D} \in \mathbb{R}^{d\times Z}$ is the learned dictionary, $\mathbf{C} \in \mathbb{R}^{Z\times N_{tr}}$ is the matrix of sparse codes, where $\mathbf{c}_j$ denotes the j-th column of $\mathbf{C}$, and Z is the dictionary size. We emphasize at this point that the sparse codes $\{\mathbf{c}_j\}_{j=1}^{N_{tr}} \in \mathbb{R}^{Z\times 1}$ are of different dimensionality compared with the sparse code vectors introduced in Section 3. However, the same symbol is used for notational convenience.

In order to enhance the discrimininative capability of the estimated sparse codes, an additional constraint is embedded in the objective function (7) as follows,

$$\hat{\mathbf{D}}, \hat{\mathbf{C}}, \hat{\mathbf{M}} = \arg\min_{\mathbf{D},\mathbf{C},\mathbf{M}} \|\mathbf{V} - \mathbf{D}\mathbf{C}\|_F^2 + \lambda_1\|\mathbf{P} - \mathbf{M}\mathbf{C}\|_F^2 , \quad \text{s.t. } \|\mathbf{c}_j\|_0 = K , \ \forall j = 1 \ldots, N_{tr} , \tag{8}$$

where λ_1 is a regularization parameter controlling the trade-off between the reconstruction error $\|\mathbf{V} - \mathbf{D}\mathbf{C}\|_F^2$ and the discriminative sparse-code error $\|\mathbf{P} - \mathbf{M}\mathbf{C}\|_F^2$. The columns of $\mathbf{P} = [\mathbf{p}_1|\cdots|\mathbf{p}_{N_{tr}}] \in \mathbb{R}^{Z\times N_{tr}}$ contain the discriminative sparse codes of the training features $\mathbf{V}$, while $\mathbf{M} \in \mathbb{R}^{Z\times Z}$ is a linear transformation matrix. In particular, $\mathbf{P}$ has a block-diagonal structure, where each one of the S blocks is an $m_i \times n_i$ matrix of ones, $\mathbf{J}_{m_i\times n_i}$, with m_i and n_i denoting the number of training feature vectors and dictionary items, respectively, which share the same class label (that is, correspond to the same speaker). In addition, $\mathbf{M}$ transforms the original sparse codes $\mathbf{C}$ so as to increase their discriminative power in the new (sparse features) space $\mathbb{R}^Z$. As a result, the discriminative sparse-code error promotes (class) label consistency in the new (transformed) sparse codes by enforcing the features from the same speaker to have similar sparse representation.

In the following, let $\mathbf{Bc}$ define a linear classifier, where $\mathbf{B} \in \mathbb{R}^{S\times Z}$ denotes the classifier parameters, and $\mathbf{c}$ is a column of the sparse code matrix $\mathbf{C}$. The output of the linear classifier will be an $S \times 1$ vector, whose largest element corresponds to the index i if the sparse code $\mathbf{c}$ is related with speaker i. Thus, in order to estimate the linear classifier parameters $\mathbf{B}$, we incorporate the classification error $\|\mathbf{H} - \mathbf{B}\mathbf{C}\|_F^2$, related with all the sparse codes contained in $\mathbf{C}$, into the objective function (8) as follows,

$$\hat{\mathbf{D}}, \hat{\mathbf{C}}, \hat{\mathbf{M}}, \hat{\mathbf{B}} = \arg\min_{\mathbf{D},\mathbf{C},\mathbf{M},\mathbf{B}} \|\mathbf{V} - \mathbf{D}\mathbf{C}\|_F^2 + \lambda_1\|\mathbf{P} - \mathbf{M}\mathbf{C}\|_F^2 + \lambda_2\|\mathbf{H} - \mathbf{B}\mathbf{C}\|_F^2 \ , \ \text{s.t. } \|\mathbf{c}_j\|_0 = K , \ \forall j = 1 \ldots, N_{tr} , \tag{9}$$

where λ_1 and λ_2 are regularization parameters controlling the trade-off between the reconstruction error $\|\mathbf{V} - \mathbf{D}\mathbf{C}\|_F^2$, the discriminative sparse-code error $\|\mathbf{P} - \mathbf{M}\mathbf{C}\|_F^2$, and the classification error $\|\mathbf{H} - \mathbf{B}\mathbf{C}\|_F^2$. Matrix $\mathbf{H} = [\mathbf{h_1}|\cdots|\mathbf{h}_{N_{tr}}] \in \mathbb{R}^{S\times N_{tr}}$ contains the class labels (or speaker index) of the training features $\mathbf{V}$. The column $\mathbf{h}_j \in \mathbb{R}^{S\times 1}$, which corresponds to the training feature vector $\mathbf{v}_j \in \mathbf{V}$ of the i-th speaker, is defined as an all-zeros vector except for the index corresponding to the true speaker label $i \in \{1, \ldots, S\}$.

The K-SVD algorithm [17] is adopted in the proposed scheme to estimate simultaneously the unknown parameters by solving the reformulated optimization problem (9) of the form

$$\hat{\mathbf{D}}, \hat{\mathbf{C}}, \hat{\mathbf{M}}, \hat{\mathbf{B}} = \arg\min_{\mathbf{D},\mathbf{C},\mathbf{M},\mathbf{B}} \left\| \begin{pmatrix} \mathbf{V} \\ \sqrt{\lambda_1}\mathbf{P} \\ \sqrt{\lambda_2}\mathbf{H} \end{pmatrix} - \begin{pmatrix} \mathbf{D} \\ \sqrt{\lambda_1}\mathbf{M} \\ \sqrt{\lambda_2}\mathbf{B} \end{pmatrix} \mathbf{C} \right\|_F^2 , \ \text{s.t. } \|\mathbf{c}_j\|_0 = K , \ \forall j = 1 \ldots, N_{tr} . \tag{10}$$

After the solution of the optimization problem (10), the estimated dictionary $\hat{\mathbf{D}}$ and classifier parameters' matrix $\hat{\mathbf{B}}$ are exploited for the final classification process. Given a test sample $\mathbf{x}_t$ we first compute its sparse representation by solving

$$\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \|\mathbf{x}_t - \hat{\mathbf{D}}\boldsymbol{\gamma}\|_2 , \ \text{s.t. } \|\boldsymbol{\gamma}\|_0 = K \tag{11}$$

through the OMP algorithm. Finally, the estimated linear classifier $\hat{\mathbf{B}}$ is applied to estimate the class (or the speaker identity) of the test sample by finding the index of the maximum value of the class label vector $\boldsymbol{\tau} = \hat{B}\hat{\gamma} \in \mathbb{R}^{S\times 1}$. As in SRC, this classification process is followed for each speech signal's frame, where finally majority voting is performed for a predefined set of frames to find the unknown speaker's identity.

5. EXPERIMENTAL RESULTS

In this section, the identification performance of the proposed discriminative K-SVD approach, described in Section 4, is evaluated in terms of the correct identification rate, and is compared with the SRC approach (discussed in Section 3) constituting the key part of the recent classification approaches for speech signals mentioned in Section 2. We also use the UBM-GMM [4] as the second method for comparison. The speech signals used in the subsequent experimental evaluations are obtained from the VOICES corpus, which is available from OGI's CSLU [18], consisting of 12 speakers (7 male and 5 female).

The original signals are sampled at 22 kHz, and downsampled to 8 kHz. During the feature extraction step, an analysis window of 320 samples, with 50% overlapping between two consecutive frames, is employed to compute a mel-frequency spectrogram of $\Omega = 40$ bands, where a silence detector algorithm based on the short-term energy and zero-crossings measure of speech segments is applied[1]. The resulting $\Omega \times T$ mel-spectrogram, where T is the total number of frames on which mel-frequency analysis was performed, is reshaped by vectorizing every ϕ consecutive columns, and thus the new matrix is of size $\phi\Omega \times \lfloor T/\phi \rfloor = \tilde{\Omega} \times \tilde{T}$. For the UBM-GMM framework a diagonal covariance matrix was chosen during the simulations. We pooled all the target speakers training data using the mel-scale frequency coefficients of order $\Omega = 40$, where after experimentation we found that best results on average obtained when used 64 number of mixtures.

[1] http://www.mathworks.com/matlabcentral/fileexchange/19298-speechcore

Table 1. Average correct identification rates (%) for the discriminative K-SVD, SRC and UBM-GMM for five different number of SNR values and four noise types: white, speech babble, car engine and factory floor. The duration of the training data is 10 sec.

Noise	SNR	K-SVD		SRC	UBM-GMM
	(dB)	25	50		
White	20	89.11	96.32	92.89	97.41
	15	86.24	97.43	87.15	98.42
	10	82.92	86.77	83.45	96.41
	5	74.75	71.96	58.71	47.70
	0	57.04	51.57	31.50	34.67
Avg.		*78.01*	*80.81*	*70.74*	*74.92*
Speech babble	20	83.95	80.41	89.88	73.90
	15	88.05	81.18	86.28	55.99
	10	80.23	83.25	70.06	30.76
	5	65.33	71.62	20.76	15.16
	0	46.43	47.55	9.46	13.77
Avg.		*72.79*	*72.80*	*55.28*	*37.91*
Engine car	20	85.52	86.45	83.29	61.55
	15	76.69	82.12	69.32	49.53
	10	50.92	64.84	65.74	34.75
	5	24.75	42.55	33.82	26.80
	0	13.55	27.65	17.36	17.85
Avg.		*50.28*	*60.72*	*53.90*	*38.09*
Factory floor	20	84.10	80.39	84.84	66.09
	15	78.32	79.92	73.16	49.39
	10	73.10	75.64	63.92	11.69
	5	45.83	59.41	16.87	8.34
	0	18.12	44.18	8.33	8.33
Avg.		*59.89*	*67.90*	*49.42*	*28.76*

It is also important to point out that for the K-SVD and SRC-based simulations $\phi = 13$ following the same vectorizing strategy as in exemplar-based techniques (ref. Section 2). In addition, $\phi = 1$ during the UBM-GMM evaluation process as a consequence of a more stable behaviour in capturing the discriminative statistics of lower dimensional features corresponding to short training data as in our study.

The duration of the training data was around 10 sec per speaker. The average correct identification rate is computed as the percentage of the correctly identified segments over the total number of test segments. For each speaker, the total number of test segments used for the evaluation is approximately equal to 70, obtained by sliding a window of 15.6 sec over the time interval of the last 10 utterances, whose duration is about 60 sec.

The test utterances are corrupted by four different types of additive noise: white noise, speech babble noise, car engine noise and factory floor noise, where the SNR of the corrupted speech takes the values of 0, 5, 10, 15 and 20 dB. The noise signals were taken from the NOISEX-92 database [19]. In all cases, the data were trained under the multicondition framework [5], where the training dataset is enlarged by corrupting the clean speech training data with simulated noise of different characteristics. Here, the clean speech data are corrupted by white noise of SNR 10, 15 and 20 dB. The sparsity threshold K mentioned in Sections 3 and 4 was chosen experimentally to be 10 during the SRC evaluation procedure, while for K-SVD a sparsity threshold equal to 25 was found to give the best performance. Besides, the regularization parameters λ_1 and λ_2 of optimization problem (10) set equal to 0.25 and 2.25 on average, respectively.

As we can see from the experimental results in Table 1, SRC achieves at least 15% higher average identification rates compared with the UBM-GMM with an exception in the case of white noise, where UBM-GMM is about 4% better. The third and fourth column correspond to the identification rates obtained using a learned K-SVD dictionary of size 25% and 50% (termed as KSVD-25 and KSVD-50) of the initial training data matrix size, respectively. It is obvious that the proposed discriminative K-SVD approach is on average far better than that of the two methods used for comparison in both dictionary size schemes. A correct identification rate of at least 60% is on average achieved with the KSVD-25 in the case of the three out of the four noise types. In addition, KSVD-50 accomplishes at least approximately 70% in three of the four noisy conditions, where in noisy conditions such as 0 and 5 dB SNR is quite robust compared with the two methods used for comparison that completely fail to achieve acceptable identification rates.

It is also important to notice how the identification rates are compared between KSVD-50 and KSVD-25. In particular, we note that KSVD-25 achieves almost similar identification rates in the case of white and speech babble noise compared to KSVD-50 and it performs lower than KSVD-50 (approximately 10% lower rates) in the case of car engine and factory floor noise. Computational cost is very crucial in real-time applications of speaker identification. In such applications we would like to achieve as high as possible correct identification rates using small amount of data. Towards this direction, KSVD-25 could be applied on 25% of the initial training data in order to achieve robust identification rates under adverse noisy conditions.

6. CONCLUSIONS

In this paper, we proposed a method for noise robust text-independent speaker identification using short training data based on a discriminative dictionary learning approach. We compare it with a UBM-GMM system, as well as with the sparse representation classification (SRC) technique. It was shown through an experimental evaluation that the proposed method performs better than the other two methods in the case of small amount of training data, and is very robust to noisy conditions. As a future work, we intend to examine the proposed approach with a larger set of realistic speech data. A further investigation could be also conducted on the theoretical part of the algorithm by introducing a non-linear classification error constraint into the objective function of (9) to achieve higher identification rates.

7. REFERENCES

[1] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. M. Chagnolleau, S. Meignier, T. Merlin, J. O. Garcia, D. P. Delacretaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430–451, 2004.

[2] J. Schmalenstroeer and R. Haeb-Umbach, "Online diarization of streaming audio-visual data for smart environments," *IEEE Selected Topics in Sig. Proc.*, vol. 4(5), pp. 845–856, October 2010.

[3] O. Vinyals and G. Friedland, "Towards semantic analysis of conversations: A system for the live identification of speak-

ers in meetings," in *Proc. Int. Conf. on Semantic Computing (ICSC)*, Santa Clara, USA, August 2008, pp. 426–431.

[4] 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.

[5] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15(5), pp. 1711–1723, July 2007.

[6] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19(7), pp. 2067–2080, September 2011.

[7] T. N. Sainath, B. Ramabhadran, M. Picheny, D. Nahamoo, and D. Kanevsky, "Exemplar-based sparse representation features: From TIMIT to LVCSR," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19(8), pp. 2598–2613, November 2011.

[8] G. S. V. S. Sivaram, S. K. Nemala, M. Elhilali, T. D. Tran, and H. Hermansky, "Sparse coding for speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, Texas, USA, March 2010, pp. 4346–4349.

[9] O. Vinyals and L. Deng, "Are sparse representations rich enough for acoustic modeling?," in *in Proc. Int. Conf. on Spoken Language Proc. (INTERSPEECH 2012)*, Portland, USA, September 2012.

[10] T. N. Sainath, D. Nahamoo, B. Ramabhadran, D. Kanevsky, V. Goel, and P. M. Shah, "Exemplar-based sparse representation phone identification features," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 4492–4495.

[11] R. Saeidi, A. Hurmalainen, T. Virtanen, and D. A. van Leeuwen, "Exemplar-based sparse representation and sparse discrimination for noise robust speaker identification," in *Proc. Odyssey: the Speaker and Language Recognition Workshop (Odyssey 2012)*, Singapore, June 2012.

[12] I. Naseem, R. Togneri, and M. Bennamoun, "Sparse representation for speaker identification," in *Proc. International Conf. on Pattern Recognition (ICPR)*, August 2010, pp. 4460–4463.

[13] B. C. Haris and R. Sinha, "Sparse representation over learned and discriminatively learned dictionaries for speaker verification," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, March 2012, pp. 4785–4788.

[14] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Maryland, USA, June 2011, pp. 1697–1704.

[15] C. Tzagkarakis and A. Mouchtaris, "Robust text-independent speaker identification using short test and training sessions," in *Proc. European Signal Processing Conf. (EUSIPCO)*, August 2010, pp. 586–590.

[16] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. on Information Theory*, vol. 53(12), pp. 4655–4666, December 2007.

[17] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Processing*, vol. 54(11), pp. 4311–4322, November 2006.

[18] A. Kain, *High Resolution Voice Transformation*, Ph.D. thesis, OGI School of Science and Engineering at Oregon Health and Science University, October 2001.

[19] A. Varga and H. J. M. Steeneken, "Assesment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247–251, 1993.