Structured Dictionary Learning for Compressive Speech Sensing

Yunyun Ji[1, 2], Wei-Ping Zhu[2] and Benoit Champagne[3]
[1] Department of Electronics and Information, Nantong University, Jiangsu, China
[2] Department of Electrical and Computer Engineering, Concordia University, Montreal, Quebec, Canada
[3] Department of Electrical and Computer Engineering, McGill University, Montreal, Quebec, Canada
Email: {jyy; weiping}@ece.concordia.ca, benoit.champagne@mcgill.ca

***Abstract*—Sparse dictionary learning aims at training appropriate redundant dictionaries for specific tasks of signal processing, such as signal estimation, compression and classification. Most of the existing dictionary learning algorithms for compressive speech sensing only exploit speech samples to construct the dictionary. In this paper, we propose to leverage both the speech signal and its linear prediction coefficients jointly to learn a structured and sparse dictionary. The proposed dictionary is designed based on a new optimization strategy using both l_0 and l_2 norms to enforce sparsity and structure, respectively. The resulting optimization problem can be solved by a fast iterative algorithm in two stages. Experimental results indicate that our proposed algorithm converges faster than the reference methods while yielding a better objective evaluation performance in terms of segmental signal-to-noise ratio, perceptual evaluation of speech quality and short-time objective intelligibility of the reconstructed speech.**

Keywords—dictionary learning; speech processing; compressive sensing; optimization

I. Introduction

Sparse representation theory plays a vital role in many fields such as speech processing, image denoising, pattern recognition, computer vision, and machine learning [1]. The core of a sparse signal representation lies in the construction of a sparsifying matrix, i.e., an orthonormal basis or overcomplete dictionary, along with the design of a sparse approximation algorithm. For the majority of applications, the specification of an appropriate sparsifying matrix determines the sparsity of the signal and further affects the overall performance of the signal processing system [2]. Compared with a simple and fixed orthonormal basis, such as the discrete cosine transform (DCT), an overcomplete dictionary learned through training can better capture the inherent structure and characteristics of the signal, yielding a sparser and more efficient representation.

The existing dictionary learning algorithms can be classified into two categories, namely: supervised and unsupervised [3]. Supervised learning aims to train the dictionary to balance the approximation error and the discrimination between various classes. Thus, supervised dictionary learning approaches are best suited for classification problems. In contrast, unsupervised dictionary learning focuses on reducing approximation errors with the constraint of sparsity. As such, the dictionary obtained by unsupervised learning is reconstructive [3] and suitable for applications such as compressive sensing (CS) [4].

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada, the National Natural Science Foundation of China (Grant No. 61601248), and the University Natural Science Research Foundation of Jiangsu Province, China (Grant No. 16KJB510037).

With large sets of training data, unsupervised dictionary learning algorithms typically solve an l_0-norm or l_1-norm minimization problem iteratively [5], [6] to generate redundant dictionaries. K-SVD [7] and online dictionary learning (ODL) [8], [9] are two representative and prevailing approaches among unsupervised dictionary learning methods. Nonetheless, in the context of speech processing, these algorithms only take advantage of the speech samples for learning the dictionary without considering the fundamental characteristics of speech signals, therefore limiting the capability of compressive speech sensing systems to exploit the signal sparsity and improve the quality of the reconstructed speech.

In this paper, by using the speech signal along with its linear prediction coefficients jointly, we propose a new optimization strategy to train a structured dictionary for speech sensing. In detail, a new structured dictionary learning (SDL) algorithm is designed to solve an original optimization problem where the l_0 and l_2 norms are employed to simultaneously enforce the sparsity and structure. Our experimental results show that our new dictionary learning algorithm outperforms the state-of-the-art reference methods K-SVD and ODL in terms of both objective evaluation of the reconstructed speech and the convergence behavior of the optimization procedure.

This paper is organized as follows: Section 2 gives a review of CS and selected unsupervised dictionary learning algorithms. In Section 3, we provide details of our proposed SDL algorithm. Section 4 presents some experimental results for the proposed method with comparison to two reference methods with respect to different objective performance measures. Finally, we conclude the paper in Section 5.

II. Compressive Sensing and Unsupervised Dictionary Learning

The essence of CS is to implement simultaneous sampling and compression of a signal at an encoder and then to reconstruct effectively the original signal from its projections at a decoder. To guarantee the sparsity of the signal, CS makes use of a redundant dictionary, represented here by a matrix $\boldsymbol{\Psi} \in \mathbb{R}^{L\times J}(L<J)$ whose columns denoted by $\boldsymbol{d}_j(j=1,2,\cdots,J)$ are referred to atoms. Let a signal vector $\boldsymbol{x} \in \mathbb{R}^L$ be K sparse with respect to a dictionary matrix $\boldsymbol{\Psi}$, i.e.,

$$\boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{s} \text{ with } \|\boldsymbol{s}\|_0 \le K \tag{1}$$

where $\boldsymbol{s} \in \mathbb{R}^J$ is the sparse coefficient vector and $\|\cdot\|_0$ denotes the l_0 norm, defined as the number of nonzero elements in a vector. Suppose that an underdetermined sensing matrix $\boldsymbol{\Phi} \in \mathbb{R}^{M\times L}$ $(M < L)$ is employed to project the signal vector $\boldsymbol{x}$ onto a lower-dimensional space. Then, the measurement vector $\boldsymbol{y} \in \mathbb{R}^M$ can be expressed as

$$\boldsymbol{y} = \boldsymbol{\Phi x} = \boldsymbol{\Phi \Psi s}. \quad (2)$$

The sensing matrix $\boldsymbol{\Phi}$ needs to satisfy the restricted isometry property (RIP) [10] to guarantee effective recovery of $\boldsymbol{x}$. Certain random matrices have been shown to fulfill this property, such as the Gaussian and Bernoulli random matrices [11]. With the above constraint on the sparsity of the coefficient vector $\boldsymbol{s}$, one can achieve effective recovery from (2) by using the basis pursuit (BP) [12] or the orthogonal matching pursuit (OMP) [13] algorithms.

Unsupervised dictionary learning algorithms have been developed to learn the dictionary by solving a sparsity-promoting minimization problem over a large set of training data. For example, the K-SVD [7] algorithm utilizes an alternating iterative method to solve an l_0-norm based sparse approximation problem,

$$\min_{\boldsymbol{\Psi},\boldsymbol{S}} \|\boldsymbol{X} - \boldsymbol{\Psi S}\|_F^2 \ \text{ s.t. } \|\boldsymbol{s}_i\|_0 \le K, \ \forall i \quad (3)$$

where $\boldsymbol{X}$ is the training data matrix, $\boldsymbol{S}$ the sparse coefficient matrix, $\boldsymbol{s}_i$ the i^{th} column vector of $\boldsymbol{S}$, and $\|\cdot\|_F$ denotes the Frobenius norm, defined as the square root of the sum of all the squared entries in a matrix. The main motivation behind the K-SVD algorithm is to employ the singular value decomposition (SVD) to iteratively update all atoms $\boldsymbol{d}_j (j = 1,2,\cdots,J)$ of the dictionary [7]. At each iteration, the representation error matrix $\boldsymbol{E}_k$ of the k^{th} atom to be updated is computed as

$$\boldsymbol{E}_k = \boldsymbol{X} - \textstyle\sum_{j\neq k} \boldsymbol{d}_j \bar{\boldsymbol{s}}_j \quad (4)$$

where $\bar{\boldsymbol{s}}_j$ refers to the j^{th} row of the sparse coefficient matrix $\boldsymbol{S}$. Then, SVD is applied to this error matrix, giving $\boldsymbol{E}_k = \boldsymbol{U\Delta V}^T$, and the atom $\boldsymbol{d}_k$ is updated by the first column of $\boldsymbol{U}$. All atoms are updated in this way at each iteration and the iteration process continues until a stopping rule is satisfied.

Alternatively, the ODL algorithm [8] is based on the following optimization problem,

$$\min_{\boldsymbol{\Psi},\boldsymbol{S}} \tfrac{1}{2}\|\boldsymbol{X} - \boldsymbol{\Psi S}\|_F^2 + \tau\|\boldsymbol{S}\|_1 \quad (5)$$

where $\|\cdot\|_1$ denotes the l_1 norm, defined as the sum of absolute values of all the entries in a matrix or a vector. The ODL algorithm exploits the fast iterative shrinkage thresholding algorithm (FISTA) to iteratively update the sparse coefficient matrix [9]. Following this step, all the atoms of the dictionary are sequentially updated at each iteration by solving the following problem,

$$\boldsymbol{\Psi} = \operatorname*{argmin}_{\boldsymbol{\Psi}} - 2\text{tr}(\boldsymbol{XS}^T\boldsymbol{\Psi}^T) + \text{tr}(\boldsymbol{SS}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}). \quad (6)$$

III. Proposed Structured Dictionary Learning

Motivated by deep learning concepts [14], the performance of a training network depends greatly upon the amount of auxiliary information fed into the network, apart from the raw data. To some extent, the dictionary can be likened to the weights of a deep neural network [14]. Thus, we may expect the dictionary to be better trained if both the raw data samples and their basic modeling features are utilized simultaneously.

One of the most efficient models to generate and approximate speech signals is the linear prediction model whose parameters, i.e. the linear prediction coefficients, play a very important role in speech coding. A p-order linear prediction model can be formulated as

$$e(n) = x(n) - \textstyle\sum_{i=1}^{p} a_i x(n-i) \quad (7)$$

where $a_i \ (i = 1,2 \cdots,p)$ and $e(n)$ represent the prediction coefficients and the prediction error individually. Due to the characteristics of speech signals, the prediction error is considered to be sparse [15]. By forming frames of L consecutive speech samples each and using (7), the speech signal for a particular frame can be written in vector form as

$$\boldsymbol{x} = \boldsymbol{\varphi e} \quad (8)$$

where $\boldsymbol{x} = [x(1) \ \cdots \ x(L)]^T$ represents the speech vector, $\boldsymbol{e} = [e(1) \ \cdots \ e(L)]^T$ is the prediction error vector, and matrix $\boldsymbol{\varphi}$ is given by

$$\boldsymbol{\varphi} = \begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots & 0 \\ -a_1 & 1 & 0 & \cdots & \cdots & 0 \\ \vdots & -a_1 & 1 & \cdots & \cdots & 0 \\ -a_p & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & -a_p & \cdots & -a_1 & 1 \end{bmatrix}^{-1}. \quad (9)$$

Matrix $\boldsymbol{\varphi}$, which is referred to as a basis in the current framework, is given in (9) as the inverse of a lower triangular matrix, and is therefore also lower triangular. Based on the sparsity of the prediction error vector $\boldsymbol{e}$ in practice, speech vector $\boldsymbol{x}$ is considered sparse with respect to the basis $\boldsymbol{\varphi}$. Moreover, the matrix $\boldsymbol{\varphi}$ is constructed from the linear prediction coefficients of speech and thus can be regarded as better capturing the internal structure of speech. Here, we propose a new optimization strategy to learn a dictionary which shall behave like the basis in (9). Unlike the existing dictionary learning algorithms which do not impart any predetermined structures on the dictionaries, our proposed dictionary $\boldsymbol{\Psi} \in \mathbb{R}^{L\times Lc}$ is structured by concatenating multiple square matrices $\boldsymbol{\psi}_j \ (j = 1,2,\cdots c) \in \mathbb{R}^{L\times L}$, i.e.,

$$\boldsymbol{\Psi} = [\boldsymbol{\psi}_1 \quad \boldsymbol{\psi}_2 \quad \cdots \quad \boldsymbol{\psi}_c] \quad (10)$$

where each submatrix $\boldsymbol{\psi}_j$ is expected, after training, to have a similar structure as $\boldsymbol{\varphi}$ in (9). As will be demonstrated in Section 4, this new structure plays a significant role in performance improvement.

Within this framework, we propose the following optimization problem for dictionary training,

$$\min_{\boldsymbol{\Psi},\boldsymbol{S}} \tfrac{1}{2}\|\boldsymbol{X} - \boldsymbol{\Psi S}\|_F^2 + \tfrac{\lambda}{2}\textstyle\sum_{j=1}^{c}\sum_{i=1}^{N}\left\|\boldsymbol{\varphi}_i - \boldsymbol{\psi}_j\right\|_F^2 \\ \text{s.t. } \|\boldsymbol{s}_i\|_0 \le K, \ \forall i \quad (11)$$

where the matrix $\boldsymbol{X} = [\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_N]$ consists of N training speech frames as its columns, the matrices $\boldsymbol{\varphi}_i$ are generated according to (9) based on the linear prediction coefficients of the i^{th} speech frame/vector in $\boldsymbol{X}$, and $\boldsymbol{S} = [\boldsymbol{s}_1 \quad \boldsymbol{s}_2 \quad \cdots \quad \boldsymbol{s}_N]$ is the sparse coefficient matrix with its i^{th} column vector $\boldsymbol{s}_i$ being the sparse coefficient vector corresponding to $\boldsymbol{x}_i$ under dictionary $\boldsymbol{\Psi}$. The parameters λ and K in (11) control the quality of the sparse approximation, the level of sparsity and the closeness of the dictionary to the true basis. Below, we propose to solve the optimization problem (11) by alternating between the sparse coding stage and dictionary updating stage.

At the sparse coding stage, with a fixed dictionary $\boldsymbol{\Psi}$, the sparse coefficient matrix is determined through solving the following problem,

$$\min_{\boldsymbol{S}} \|\boldsymbol{X} - \boldsymbol{\Psi}\boldsymbol{S}\|_F^2 \text{ s.t. } \|\boldsymbol{s}_i\|_0 \leq K, \ \forall i. \qquad (12)$$

This optimization problem can be further divided into N distinct subproblems with respect to the sparse coefficient vectors, i.e. the column vectors of the sparse matrix $\boldsymbol{S}$,

$$\min_{\boldsymbol{s}_i} \|\boldsymbol{x}_i - \boldsymbol{\Psi}\boldsymbol{s}_i\|_2^2 \text{ s.t. } \|\boldsymbol{s}_i\|_0 \leq K, \quad \text{for} \quad i = 1,2, \cdots N. \qquad (13)$$

Even though a typical l_0-norm based minimization problem is NP-hard, we can utilize a greedy pursuit algorithm like OMP as an alternative to estimate the sparse coefficient vectors.

At the dictionary updating stage, with the sparse coefficient matrix $\boldsymbol{S}$ obtained from the sparse coding stage, the dictionary $\boldsymbol{\Psi}$ can be updated by solving the following problem,

$$\min_{\boldsymbol{\Psi}} \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{\Psi}\boldsymbol{S}\|_F^2 + \frac{\lambda}{2}\sum_{j=1}^{c}\sum_{i=1}^{N}\left\|\boldsymbol{\varphi}_i - \boldsymbol{\psi}_j\right\|_F^2. \qquad (14)$$

In this paper, we propose to update the dictionary $\boldsymbol{\Psi}$ block by block, unlike the traditional atom-by-atom update as in K-SVD and ODL. The coefficient matrix $\boldsymbol{S}$ can be partitioned per rows according to the order of the bases in $\boldsymbol{\Psi}$, i.e., $\boldsymbol{S} = [\boldsymbol{\theta}_1^T \quad \boldsymbol{\theta}_2^T \quad \cdots \quad \boldsymbol{\theta}_c^T]^T$. In this case, the gradient of (14) with respect to $\boldsymbol{\psi}_j$ can be written as $-(\boldsymbol{X} - \boldsymbol{\Psi}\boldsymbol{S})\boldsymbol{\theta}_j^T + \lambda\sum_{i=1}^{N}(\boldsymbol{\psi}_j - \boldsymbol{\varphi}_i)$. By letting it be zero, the submatrix $\boldsymbol{\psi}_j$ in the dictionary can be updated as

$$\boldsymbol{\psi}_j = \left((\boldsymbol{X} - \textstyle\sum_{k\neq j}\boldsymbol{\psi}_k\boldsymbol{\theta}_k)\boldsymbol{\theta}_j^T + \lambda\sum_{i=1}^{N}\boldsymbol{\varphi}_i\right)\left(\boldsymbol{\theta}_j\boldsymbol{\theta}_j^T + \lambda N\boldsymbol{I}\right)^{-1} \qquad (15)$$

where $\boldsymbol{I} \in \mathbb{R}^{L\times L}$ is an identity matrix. The pseudo-code of our proposed SDL algorithm is presented in Algorithm 1.

Algorithm 1: Structured Dictionary Learning Algorithm

Input: data matrix $\boldsymbol{X}$, true bases $\boldsymbol{\varphi}_i (i = 1,2, \cdots N)$, λ, τ

Initialization: randomly select c true bases and fill them into (10) to generate the initial dictionary $\boldsymbol{\Psi}^0$.

Iteration: at the l^{th} iteration,

Sparse coding stage: with the dictionary $\boldsymbol{\Psi}^{l-1}$, OMP is employed to solve the following optimization problem to update the sparse coefficient matrix,

$$\boldsymbol{S}^l = \underset{\boldsymbol{S}}{\operatorname{argmin}} \left\|\boldsymbol{X} - \boldsymbol{\Psi}^{l-1}\boldsymbol{S}\right\|_F^2 \text{ s.t. } \|\boldsymbol{s}_i\|_0 \leq K, \ \forall i.$$

Dictionary updating stage: the sub-matrices in the dictionary are updated sequentially. The residual of the data excluding the j^{th} square matrix can be calculated as

$$\boldsymbol{R}_j^l = \boldsymbol{X} - \textstyle\sum_{k\neq j}\boldsymbol{\psi}_k^{l-1}\boldsymbol{\theta}_k^l$$

and the j^{th} square matrix is updated as

$$\boldsymbol{\psi}_j^l = \left(\boldsymbol{R}_j^l\boldsymbol{\theta}_j^{l^T} + \lambda\textstyle\sum_{i=1}^{N}\boldsymbol{\varphi}_i\right)\left(\boldsymbol{\theta}_j^l\boldsymbol{\theta}_j^{l^T} + \lambda N\boldsymbol{I}\right)^{-1}.$$

The dictionary is obtained as $\boldsymbol{\Psi}^l = [\boldsymbol{\psi}_1^l \quad \boldsymbol{\psi}_2^l \quad \cdots \quad \boldsymbol{\psi}_c^l]$. The above iteration continues until a stopping criterion is satisfied.

Output: Sparse coefficient matrix $\boldsymbol{S}$ and dictionary $\boldsymbol{\Psi}$.

IV. Experimental Results

In this section, we evaluate the reconstruction performance of our proposed SDL algorithm by using the speech corpus GRID [16]. GRID is a multi-talker corpus, consisting of speech utterances from 18 males and 16 females with the maximum signal amplitude normalized to 1. We randomly select 5 male and 5 female speakers from this corpus and downsample the utterances to 16 KHz. For each speaker, 70 sentences are selected randomly as training data, while another 10 sentences distinct from the training set are used as the test set. Due to lack of space, in this paper, we only present the results for the speaker dependent case, i.e. where the test dataset and training dataset are from the same speakers. However, the results for the speaker-independent case show similar improvements in comparison to the benchmark methods like K-SVD and ODL.

We employ segmental signal-to-noise ratio (SSNR) [17], perceptual evaluation of speech quality (PESQ) [18], and short-time objective intelligibility (STOI) score [19] as objective measures to evaluate the quality of the recovered speech. We also compare our method with two state-of-the-art dictionary learning algorithms, namely, K-SVD [7] and ODL [8]. The parameters used in our experiments are set as follows: frame length $L = 512$, linear prediction model order $p = 8$, regulation factor $\lambda = 0.01$, sparsity level $K = 5$ and dictionary size $c = 3$ (i.e., the dictionary size is three times the frame size). The simulation results are averaged over all the testing data of male and female speakers. In what follows, we show the convergence behavior and the performance evaluation results of the proposed SDL algorithm with comparison to K-SVD and ODL. We consider two cases, that is, the noise-free case where the observations are not corrupted by noise, and the noisy case in which observations are corrupted by zero-mean additive white Gaussian noise (AWGN). In the noise-free case, the BP algorithm is employed to reconstruct the speech signals using the dictionaries returned by SDL, K-SVD and ODL.

Table I presents the average PESQ of the reconstructed speech in the noise-free case using different dictionaries for both male and female speakers at various compression rates. It is obvious that our structured dictionary achieves better average PESQ scores than the reference methods. For instance, when $M/L = 0.5$, the average PESQ score for female speakers is increased from 2.47 in K-SVD and 2.49 in ODL to 2.74 in the proposed SDL algorithm. Fig. 1 demonstrates the average SSNR of the reconstructed speech at different compression rates for male speakers (a) and female speakers (b) based on different dictionaries. The average SSNRs of the SDL algorithm are larger than those of the K-SVD and ODL at all the compression rates. For example, when the compression rate is 0.5, the average SSNR of male speakers with our proposed dictionary is 20.54 dB which is about 5 dB higher than that resulting from K-SVD and ODL, clearly showing the advantage of SDL in speech coding under noise-free condition.

ISBN 978-90-827970-1-5 © EURASIP 2018

TABLE I. AVERAGE PESQ RESULTS IN NOISE-FREE ENVIRONMENT

M/L	Male Speaker			Female Speaker		
	K-SVD	ODL	SDL	K-SVD	ODL	SDL
0.3	2.13	2.22	**2.35**	2.17	**2.22**	**2.22**
0.4	2.28	2.38	**2.57**	2.32	2.35	**2.50**
0.5	2.43	2.51	**2.78**	2.47	2.49	**2.74**
0.6	2.60	2.67	**2.94**	2.63	2.65	**2.81**
0.7	2.76	2.82	**3.10**	2.81	2.81	**3.18**
0.8	2.96	2.97	**3.27**	2.99	3.00	**3.39**

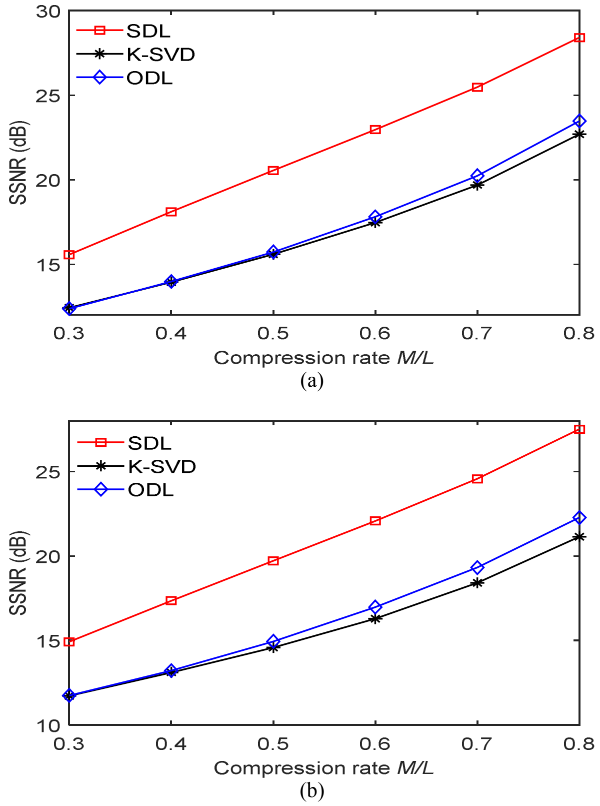

Fig. 1. Average SSNR results in noise-free environment: (a) Male speakers; (b) Female speakers.

It is also worth-considering the robustness of the proposed SDL algorithm in the noisy environment. In this setting, the standard deviation of the AWGN is 0.002. Here, we have employed the basis pursuit denoising (BPDN) algorithm [12] as the recovery method. Table II and Fig. 2 give the average PESQ scores and average SSNR results using the different dictionaries in the noisy environment. Again, the results show the proposed SDL based dictionary gives better performance than the other two algorithms. For instance, when $M/L = 0.5$, the average PESQ scores of our proposed SDL algorithm are 2.60 and 2.55 for male and female speakers. At the same compression rate, K-SVD can achieve the PESQ scores of 2.39 and 2.42 respectively for both genders, while those of ODL are 2.47 and 2.51 respectively. Moreover, under the same condition, the average SSNRs of the SDL algorithm are respectively around 5dB and 4dB higher than that of K-SVD and ODL for male speakers and female speakers. From Fig. 3, we can also see that the average STOI scores of the reconstructed speech signals from the SDL-based dictionary are higher than those of the other methods, demonstrating that the SDL dictionary can lead to a better intelligibility of the reconstructed speech. For example, in Fig. 3(a), the improvement of our proposed SDL algorithm in STOI is 0.07 and 0.04 respectively over K-SVD and ODL at the compression rate of 0.5 for male speakers. One can also conclude that the SDL algorithm is more robust than the K-SVD and ODL algorithms in noisy environment. Moreover, in informal listening tests, we find that the speech files processed with the proposed SDL method exhibit improved quality and more naturalness than those processed with reference methods.

Similar to [20], we evaluate the convergence behavior of our proposed algorithm via the increment of the sparse coefficient matrix sequences $\{\boldsymbol{S}^l\}$, defined as $\|\boldsymbol{S}^{l+1} - \boldsymbol{S}^l\|_F$ versus the iteration number. As stated in [7], convergence of the K-SVD algorithm cannot be guaranteed; hence, we compare only our algorithm with ODL which is proved to be convergent in [8]. The average values of the sequence $\{\|\boldsymbol{S}^{l+1} - \boldsymbol{S}^l\|_F\}$ for the two algorithms versus the iteration number are plotted in Fig. 4. In terms of this metric, the proposed SDL algorithm converges faster than the ODL algorithm.

TABLE II. AVERAGE PESQ RESULTS IN NOISY ENVIRONMENT

M/L	Male Speaker			Female Speaker		
	K-SVD	ODL	SDL	K-SVD	ODL	SDL
0.3	2.25	2.34	**2.38**	2.26	**2.37**	2.27
0.4	2.33	2.42	**2.50**	2.36	**2.45**	**2.45**
0.5	2.39	2.47	**2.60**	2.42	2.51	**2.55**
0.6	2.43	2.51	**2.66**	2.48	2.56	**2.64**
0.7	2.46	2.53	**2.70**	2.51	2.59	**2.70**
0.8	2.49	2.55	**2.74**	2.54	2.61	**2.76**

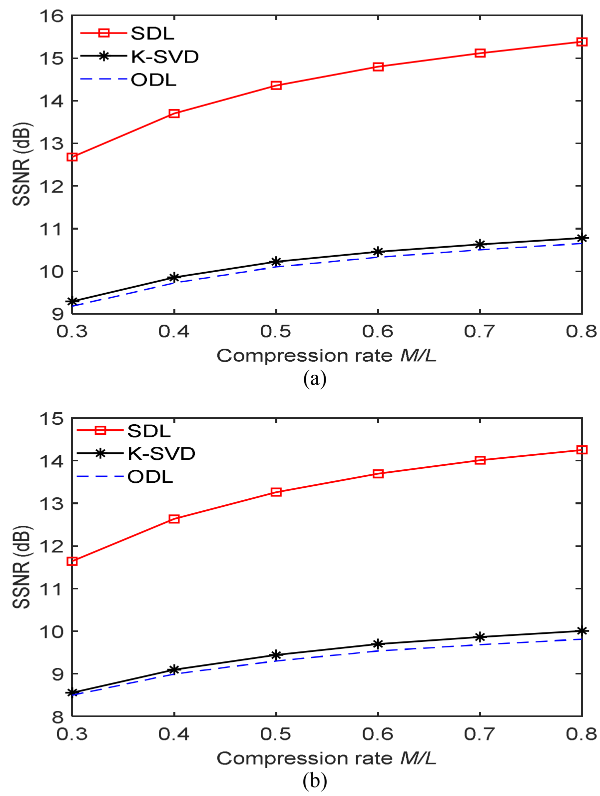

Fig. 2. Average SSNR results in noisy environment: (a) Male speakers; (b) Female speakers.

ISBN 978-90-827970-1-5 © EURASIP 2018

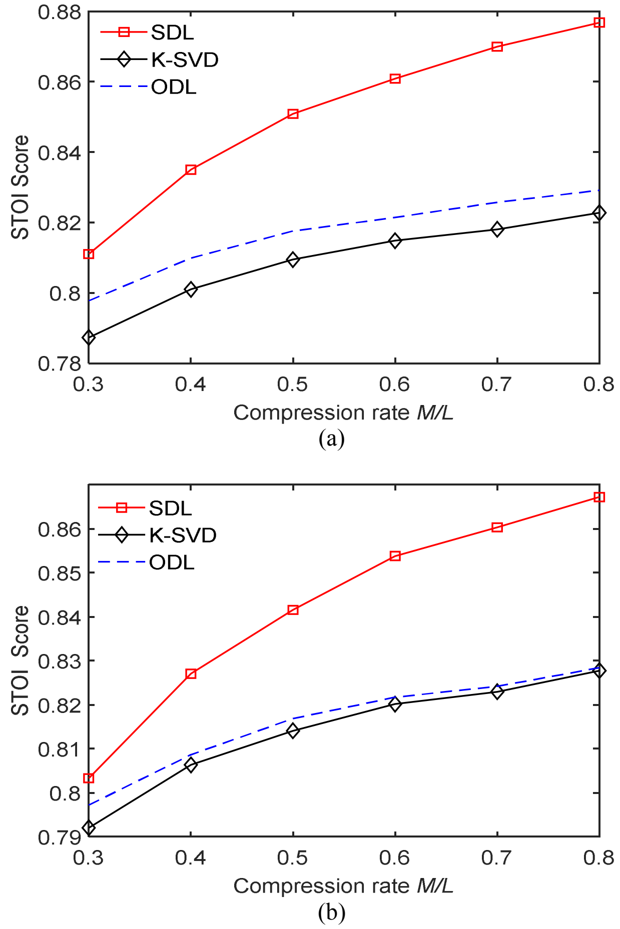

Fig. 3. Average STOI scores in noisy environment: (a) Male speakers; (b) Female speaker.

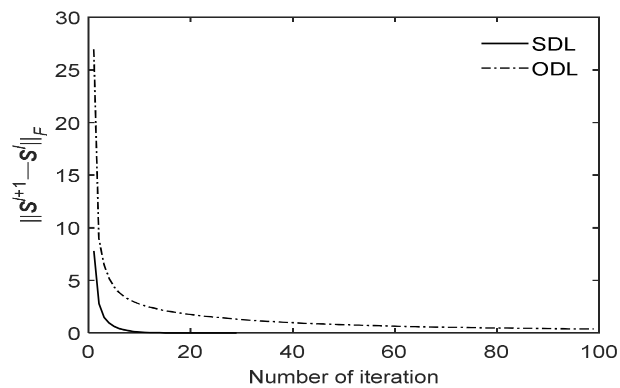

Fig. 4. Convergence behavior of SDL and ODL algorithms: increment norm of sparse coefficients matrix versus iteration number

V. Conclusion

In this paper, we have proposed a structured dictionary learning algorithm for compressive speech sensing by formulating and solving a novel optimization problem. In contrast to the existing dictionary learning methods, our dictionary is constructed by taking advantage of the linear prediction model and a special structure. It is trained with both speech samples and corresponding linear prediction coefficients as known features. It is shown through extensive experimental studies that the proposed dictionary learning method yields a higher quality of reconstructed speech in both noise-free and noisy conditions and a fast convergence as compared to two state-of-the-art methods.

References

[1] R. Rubinstein, A. M. Bruckstein, and M. Elad, “Dictionaries for sparse representation modelling,” *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045-1057, June 2010.

[2] Z. Zhang, Y. Xu, J. Yang, X. L. Li and D. Zhang, “A survey of sparse representation: algorithms and applications,” *IEEE Access*, vol. 3, pp. 490-530, May 2015.

[3] I. Tosic and P. Frossard, “Dictionary learning,” *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27-38, March 2011.

[4] S. Foucart and H. Rauhut, A Mathematical Introduction to Compressive Sensing, Borkhauser, New York, 2013.

[5] M. Elad, Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing, Springer, NY, 2010.

[6] A. M. Bruckstein, D. L. Donoho and M. Elad, “From sparse solutions of systems of equations to sparse modeling of signals and images,” *SIAM Review*, vol. 51, pp.34-81, 2009.

[7] M. Aharon, M. Elad and A. Brukstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE Trans. on Signal Processing,* vol. 54, no.11, November 2006.

[8] J. Mairal, F. Bach, J. Ponce and G. Sapiro, “Online learning for matrix factorization and sparse coding,” *The Journal of Machine Learning Research*, vol. 11, pp. 19-60, January 2010.

[9] T. H. Vu and V. Monga, “Fast low-rank shared dictionary learning for image classification,” *IEEE Trans. on Image Processing*, vol. 26, no. 11, pp. 5160-5175, November 2017.

[10] E. J. Candes, J. Romberg and T. Tao, “Stable signal recovery from incomplete and inaccurate measurements,” *Communications on Pure Applied Mathematics*, vol. 59, no.8, pp. 1207-1223, March 2006.

[11] R. Baraniuk, M. Davenport, R. Devore and M. Wakin, “A simple proof of the restricted isometry property for random matrices,” *Constructive Approximation*, vol. 28, pp. 253-263, January 2008.

[12] S. S. Chen, D. L. Donoho and M. A. Saunders, “Atomic Decomposition by Basis Pursuit,” *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33-61, July 2006.

[13] J. Tropp and A. C. Gilbert, “Signal recovery from random measurement via orthogonal matching pursuit,” *IEEE Trans. on Information Theory*, vol. 53, no. 12, pp.4655-4666, 2007.

[14] Y. Xu, J. Du, L. R. Dai and C. H. Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 7-19, Jan. 2015.

[15] D. Giacobello, M. D. Christensen, M. N. Murthi, S. H. Jensen and M. Moonen, “Retrieving sparse pattern using a compressed sensing framework: application to speech coding based on sparse linear prediction,” *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 103-106, October 2009.

[16] C. D. Sigg, T. Dikk, and J. M. Buhmann, “Speech enhancement using generative dictionary learning,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 20, no. 6, pp.1698-1712, August 2012.

[17] J. Du and Q. Huo, “A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortion,” *in Proc. Interspeech*, pp.569-572, January 2008.

[18] Y. Hu and P. C. Loizou, “Evaluation of objective quality measure for speech enhancement,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 229-238, January 2008.

[19] C. H. Taal, R. C. Hensick, R. Heusdens and J. Jensens, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125-2136, September 2011.

[20] C. L. Bao, H. Ji, Y. H. Quan, and Z. W. Shen, “Dictionary learning for sparse coding: algorithms and convergence analysis,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 38, no. 7, pp. 1356-1369, July 2016.

ISBN 978-90-827970-1-5