DESIGN OF OPTIMAL MATRICES FOR COMPRESSIVE SENSING: APPLICATION TO ENVIRONMENTAL SOUNDS

Bochra BOUCHHIMA [1], *Rim AMARA* [1,2], *Monia Turki-Hadj Alouane*[1]

[1] Université de Tunis El Manar,
Ecole Nationale d'Ingénieurs de Tunis,
$03/\mathrm{UR}/11-07$
Laboratoire de Signaux et systémes,
1002, Tunis, Tunisie

[2] Université de Carthage,
Institut National des Sciences
Appliquées et de Technologie
1080, Tunis, Tunisie

ABSTRACT

In a compressive sensing context, we propose a solution for a full learning of the dictionary composed of the sparsity basis and the measurement matrix. The sparsity basis learning process is achieved using Empirical Mode Decomposition (EMD) and Hilbert transformation. EMD being a data-driven decomposition method, the resulting sparsity basis shows high sparsifying capacities. On the other hand, a gradient method is applied for the design of the measurement matrix. The method integrates the dictionary normalization into the target function. It is shown to support large scale problems and to have a good convergence and high performance. The evaluation of the whole approach is done on a set of environmental sounds, and is based on a couple of key criteria: sparsity degree and incoherence. Experimental results demonstrate that our approach achieves well with regards to mutual coherence reduction and signal reconstruction at low sparsity degrees.

Index Terms— Compressive Sensing, EMD, Environmental Sounds, Sparsity, Measurement Matrix, Incoherence

1. INTRODUCTION

Compressive sensing (CS) is a quite recent paradigm applied for signal and image processing. It jointly acquires and compresses signals without going through intermediate processing stages as in the traditional acquisition systems. This concept is made possible given that most signals are compressible or sparse in some convenient basis. Formally, consider $\boldsymbol{x} \in \mathbf{R}^N$. $\boldsymbol{x}$ is said to be sparse if there exists a basis $\boldsymbol{\Psi} \in \mathbf{R}^{N \times K}$ in which it is expressed as $\boldsymbol{x} = \boldsymbol{\Psi s}$ with $\boldsymbol{s}$ having at most $P \ll N$ non zeros elements. $\boldsymbol{\Psi}$ is called the sparsity basis. Now, consider $\boldsymbol{x}$ as the signal to be acquired in a wireless sensor network (WSN) transmission for tele-monitoring context. From an energy perspective, it is recommended to reduce operations complexity at each sensor node so as to prolong its lifetime. The classical sampling methods would require N samples of $\boldsymbol{x}$ and a compression stage afterwards. Compressive sensing, however, suggests to acquire only $M < N$ samples given by the linear multiplication of $\boldsymbol{x}$ by a measurement matrix $\boldsymbol{\Phi} \in \mathbf{R}^{M \times N}$. Let $\boldsymbol{y} = \boldsymbol{\Phi\Psi s} = \boldsymbol{Ds}$ be the measurement vector. $\boldsymbol{x}$ can be reconstructed from $\boldsymbol{y}$ by running optimisation algorithms [1]. Applying a successful CS scenario for signals consists therefore in two major tasks : designing "good" matrices both for sparsity and measurements, and designing a robust algorithm to recover $\boldsymbol{s}$ from $\boldsymbol{y}$.

Problem formulation

As far as this work is concerned, we focus on the learning of a sparsity matrix and a measurement matrix responding to a couple of key notions: sparsity and incoherence. The first notion is related to $\boldsymbol{\Psi}$. $\boldsymbol{\Psi}$ is said to be a convenient matrix if signal recovery can be achieved from a sparse form $\boldsymbol{s}$ having a low sparsity degree P. The second notion, incoherence, stands for the relation between $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ and measures how uncorrelated $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are. When both matrices are incoherent, there is an overwhelming probability to achieve a good signal recovery. In this study, we propose an approach for designing a sparsity matrix with high sparsity capacities and a measurement matrix incoherent to it. Both are ensuring a good signal reconstruction.

State of the art

It has been established that learning adaptive basis from a set of training signals yields better results than using a predefined basis. Many Dictionary learning algorithms were therefore implemented such as MOD [2], KSVD [3], GAD [4], etc. In a previous work [5], we proposed a new learning approach based on EMD and Hilbert transformation. Once the sparsity basis $\boldsymbol{\Psi}$ is known, one particular rule should be respected to define a measurement matrix $\boldsymbol{\Phi}$: both matrices, $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ must be incoherent. The majority of works made the assumption that $\boldsymbol{\Phi}$ is drawn from a gaussian distribution of variance $1/M$. Yet, learning a projection matrix can indeed reduce the mutual coherence and the reconstruction error as well. This was proved in [6–10]. [6] used a technique of shrinkage to optimize a so-called "t-averaged mutual coherence". [7] in-

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

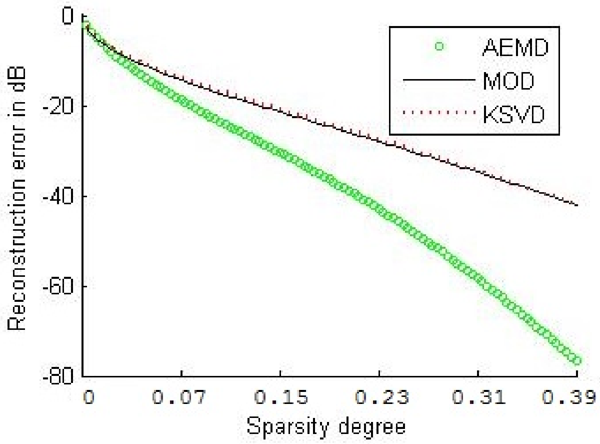

Fig. 1. Reconstruction error averaged over 300 test signals while using matrices resulting from AEMD, KSVD and MOD. Sparsity degree stands for the ratio $= \frac{number\ of\ non\ null\ elements}{signal\ size}$. For higher sparsity degrees, refer to [5].

troduces a framework for simultaneously designing $\mathbf{\Phi}$ and $\mathbf{\Psi}$ and applies SVD decomposition to solve the corresponding problem. [8] suggests to deal with a constrained problem and solve it using the method of alternating projections. [9] suggests to design an equiangular tight frame and uses QR factorization to define $\mathbf{\Phi}$. In [10], Abolghasemi et al. use the gradient descent method for learning the measurement matrix. At each iteration, they proceed to the normalization of the dictionary atoms first.then, they apply the gradient descent update. Their approach is interesting in terms of low complexity. Yet, its convergence is problematic and its performance can be enhanced. In this paper, we propose an EMD-based sparsity matrix learning, followed by a gradient-based measurement matrix learning method which integrates the dictionary normalization operation. The first part focuses on achieving a good sparsity on a set of environmental sounds. While the second part attempts to reduce the mutual coherence. The remainder of this paper is organised as follows: Section 2 reminds EMD principle and describes our method for sparsity matrix learning. In section 3, we formally address the problem of incoherence and describe our method for learning $\mathbf{\Phi}$. Section 4 displays the results of simulations. Finally, we come up with a conclusion to sum up the main ideas and results.

2. SPARSITY MATRIX LEARNING

The success of compressive sensing relies, in part, on the design of the sparsity matrix which should fit the structure of the signals. In [5], we used EMD to learn an over-complete sparsity matrix in a non-iterative way.

2.1. Empirical mode decomposition

EMD is an adaptive decomposition proposed by N. E. Huang in 1998 [11]. It states that a signal $\boldsymbol{x}$ can be written as a sum of some components called intrinsic mode functions (IMF) and a trend (T) as follows

$$\boldsymbol{x} = (\sum_{i \leq m} \boldsymbol{IMF^i}) + \boldsymbol{T} \quad (1)$$

Each IMF is a simple oscillatory mode in a frequency limited-band (see [11] for details about the sifting process). Theoretically, IMFs of a given signal are mutually orthogonal. Yet, in practice, they are often close to be orthogonal. In particular, when the signal is narrow-band, orthogonality is not reached.

2.2. Analytic EMD for sparsity matrix learning

We denote AEMD (Analytic Empirical Mode Decomposition) the process explained in algorithm 1. Provided a learning set, we apply EMD for each signal. The IMFs to be kept are selected according to their relevance to the signal. We measure the relevance of each IMF by introducing the Signal-to-IMF-Ratio [5] and we find out that the first three IMFs are the most important, the rest can be considered as trend. Next, we apply Hilbert transformation to transform the IMFs into analytic form before proceeding to a clustering phase using K-means algorithm. Here, we slightly modify the algorithm proposed in [5] so as to reduce the dictionary size and the mutual coherence between $\mathbf{\Phi}$ and $\mathbf{\Psi}$ as well. In fact, the trend resulting from the signals decomposition is no more included in the learning process. We find out experimentally that the mutual coherence is more probably to get reduced when sparsity matrix does not contain clusters of trend signals. We show in Fig.1 that our method outperforms algorithms used in [3] and [2]. We focus on low sparsity degrees, as recommended in CS context.

Algorithm 1: AEMD

Input: Learning data set $\{\boldsymbol{x}_i\}_{i=1}^{L}$
Initialization: $\mathbf{\Psi}^j_{Re} \leftarrow [\,]$; $\mathbf{\Psi}^j_{An} \leftarrow [\,]$; $1 \leq j \leq 3$
for $i = 1$ **to** L
  $[\boldsymbol{IMF}^1_i, \boldsymbol{IMF}^2_i, \boldsymbol{IMF}^3_i, \boldsymbol{T}_i] \leftarrow EMD(\boldsymbol{x}_i)$
  for $j = 1$ **to** 3
    $\mathbf{\Psi}^j_{Re} \leftarrow [\mathbf{\Psi}^j_{Re} \mid \boldsymbol{IMF}^j_i]$ # IMFs extracted at a level j are stacked along the atoms of $\mathbf{\Psi}^j_{Re}$
  end for
end for
for $j = 1$ **to** 3
  $\mathbf{\Psi}^j_{An} \leftarrow [HT(\mathbf{\Psi}^j_{Re})]$ #Hilbert Transformation of the atoms of each sub$-$dictionary
  $\{\mathbf{\Psi}^j\} \leftarrow$ k$-$means($\mathbf{\Psi}^j_{An}, C^j$)# create C^j clusters at each level
end for
$\mathbf{\Psi} = [\mathbf{\Psi}^1, \mathbf{\Psi}^2, \mathbf{\Psi}^3]$
Normalize $\mathbf{\Psi}$ in columnwise sense

3. MEASUREMENTI MATRIX LEARNING

Incoherence (μ) has a major influence on the theoretical performance of signals' reconstruction. Commonly, the incoherence is mathematically formulated as

$$\mu_{max} = \max_{1\leq i,j\leq K, i\neq j} \mid \boldsymbol{G}_{i,j} \mid = \max_{1\leq i,j\leq K, i\neq j} \mid \widetilde{\boldsymbol{d}}_i^H \widetilde{\boldsymbol{d}}_j \mid \quad (2)$$

where $\widetilde{\boldsymbol{d}}_i$ denotes the i^{th} column of $\boldsymbol{D}$ scaled to unity (in l_2 norm sense), $\boldsymbol{G}$ is the gram matrix and $.^H$ is the transpose conjugate. Note that $0 < \mu \leq 1$. Two matrices are incoherent if μ approaches 0 .Yet, considering only the maximum value as an optimization criteria may not be relevant enough. Consider the case in which the gram matrix has one singular off-diagonal value which is close to 1 while all the other off-diagonal components are very close to zero. According to the definition of μ_{max}, the mutual coherence will be judged as unsatisfactory and the two matrices as coherent. In order to mitigate this situation, [10] consider the averaged incoherence defined as

$$\mu_{avg} = \frac{\sum_{1\leq i,j\leq K, i\neq j} \mid \widetilde{\boldsymbol{d}}_i^H \widetilde{\boldsymbol{d}}_j \mid}{K(K-1)} \quad (3)$$

So far, we will mainly focus on the μ_{avg} values in evaluating the performance of our method.

3.1. Gradient approach

Let $\widetilde{\boldsymbol{D}}$ denote the column-normalized version of $\boldsymbol{D}$. To achieve incoherence having $\boldsymbol{\Psi}$ furnished by AEMD, we are looking to find $\boldsymbol{\Phi}$ such that for $\boldsymbol{D} = \boldsymbol{\Phi\Psi}$, we have

$$\widetilde{\boldsymbol{D}}^H \widetilde{\boldsymbol{D}} \simeq I_{K\times K} \quad (4)$$

It is easy to see that

$$\widetilde{\boldsymbol{D}} = \boldsymbol{\Phi}\widetilde{\boldsymbol{\Psi}} = \boldsymbol{\Phi\Psi S} \quad (5)$$

where $\boldsymbol{S} = diag(\frac{1}{\|\boldsymbol{\Phi\Psi}_1\|_2}, \frac{1}{\|\boldsymbol{\Phi\Psi}_2\|_2}, ..., \frac{1}{\|\boldsymbol{\Phi\Psi}_K\|_2})$.

$\boldsymbol{\Psi}_i$ is the i^{th} column of $\boldsymbol{\Psi}$. $\boldsymbol{S}$ is well defined because $\forall i, \|\boldsymbol{\Phi\Psi}_i\|_2 \neq 0$.
Equation (4) becomes

$$\boldsymbol{S}^H\boldsymbol{\Psi}^H\boldsymbol{\Phi}^H\boldsymbol{\Phi\Psi S} \simeq I_{K\times K} \quad (6)$$

$\boldsymbol{S}$ is hermitian and invertible. By multiplying both sides by $\boldsymbol{S}^{-1}$ we get

$$\boldsymbol{\Psi}^H\boldsymbol{\Phi}^H\boldsymbol{\Phi\Psi} \simeq (S^{-1})^2 \quad (7)$$

7can then be solved through the Least square (LS) minimization problem

$$\widehat{\boldsymbol{\Phi}} = \min_{\boldsymbol{\Phi}} \; \mathrm{J}(\boldsymbol{\Phi}) = \left\| \boldsymbol{\Psi}^H\boldsymbol{\Phi}^H\boldsymbol{\Phi\Psi} - (S^{-1})^2 \right\|_F^2 \quad (8)$$

We use the Frobenius norm $\|.\|_F$ to get a differentiable cost function. Moreover the choice of this norm is convenient for our optimization problem. In fact, $\|.\|_F$ optimizes the entries square which is stronger than optimising the absolute value as suggested by (3). To solve (8), we choose to apply the gradient descent algorithm which is an iterative method aiming at finding the nearest local minimum of a function starting at an initial point $\boldsymbol{\Phi}_0$ and moving from $\boldsymbol{\Phi}_i$ to $\boldsymbol{\Phi}_{i+1}$ in the opposite direction of the gradient. Let's start by defining the derivative of the cost function

$$\mathrm{J}(\boldsymbol{\Phi}) = Tr(\boldsymbol{\Psi}^H\boldsymbol{\Phi}^H\boldsymbol{\Phi\Psi} - (S^{-1})^2)(\boldsymbol{\Psi}^H\boldsymbol{\Phi}^H\boldsymbol{\Phi\Psi} - (S^{-1})^2)^H \quad (9)$$

Without going into calculation details, we can show that the gradient is

$$\nabla \mathrm{J} = 4\boldsymbol{\Phi\Psi}(\boldsymbol{\Psi}^H\boldsymbol{\Phi}^H\boldsymbol{\Phi\Psi} - (S^{-1})^2)\boldsymbol{\Psi}^H \quad (10)$$

$\boldsymbol{\Phi}$ is updated then at each iteration as follows

$$\boldsymbol{\Phi}_{i+1} = \boldsymbol{\Phi}_i - \delta \, \nabla \mathrm{J}(\boldsymbol{\Phi}_i) \quad (11)$$

where δ is the stepsize. The gradient descent algorithm is also sensitive to the choice of the step. A small step value makes the convergence slower while a big stepsize may cause the algorithm to diverge.

4. EXPERIMENTS AND RESULTS

In all experiments, we compare our method to that of [10], which was shown to outperform the method in [6]. We label the approach in [10] by "2-stage gradient method" in contrast with ours which is "1-stage gradient method". In fact, at each iteration, [10] gets the dictionary normalized before updating the $\boldsymbol{\Phi}$. In our contribution, and as we claim before, we integrate the normalization factor into the cost function via the matrix $\boldsymbol{S}$ in (6). So far, the experiments are carried out on short-duration, also called environmental or impulsive, sounds related to tele$-$monitoring context. Some are expressing emergency situations (screams, glass breaking), while the others are usual sounds (cat mewing, door slamming,...). They are issued from various sound libraries available on the net. The sampling frequency is 11 025 Hz and all signals are mono-channel. The learning set contains 800 patterns of $N = 256$ coefficients length each. In fact, to maintain a fixed length for all patterns, we split audio signals into frames of the same length (256 samples). The sparsity matrix is composed of $K = 750$ atoms and is learned as explained in algorithm 1. We take $M = 120$. In order to observe the cost function variation for both methods, many experiments have been conducted while varying step sizes. Afterwards, we select the step size giving the best convergence properties for each method: 4.10^{-6} for the 2-stage gradient method and 4.10^{-5} for ours. Fig.2 shows the cost function variation for different initialisations ($I_{M\times N}$ is the identity matrix of size $M \times N$, $\mathrm{U}_{[0,1]}$ is the uniformly distributed entries matrix, $\mathcal{N}(0, \frac{1}{M})$ is an i.i.d distributed matrix with variance $\frac{1}{M}$, Al_1

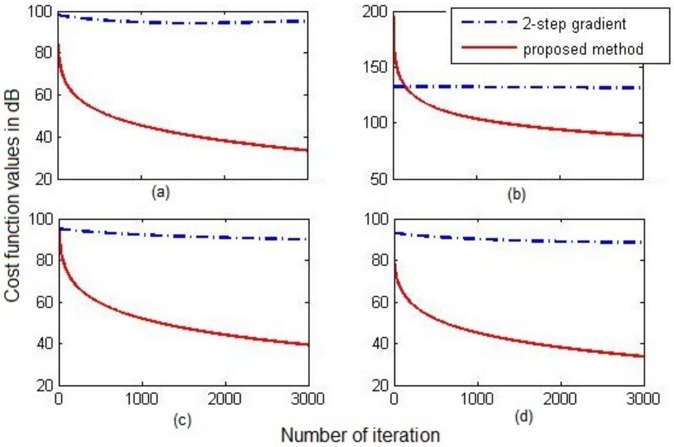

Fig. 2. Convergence of our method (continuous line) versus the 2-step gradient method (dotted line) for different initialisation matrices $\mathbf{\Phi}_0$ and $\mathbf{\Psi}$ EMD-learned matrix. (a)$\mathbf{\Phi}_0 = I_{M\times N}$, (b)$\mathbf{\Phi}_0 \sim \mathrm{U}_{[0,1]}$, (c)$\mathbf{\Phi}_0 \sim \mathcal{N}(0, \frac{1}{M})$, (d)$\mathbf{\Phi}_0 = Al_1$

$\mathbf{\Phi}_0$	$I_{M\times N}$	$\mathrm{U}_{[0,1]}$	$\mathcal{N}(0, \frac{1}{M})$	Al_1
Initial values	0.51	0.41	0.46	0.41
Proposed approach	0.11	0.10	0.09	0.08
2-stage gradient	0.11	0.25	0.10	0.09

Table 1. μ_{avg} values returned by both algorithms for different initialisations.

is a matrix containing only 1 valued entry per line in a random way, the remaining components are all set to 0). Note that at each iteration, there is $120 \times 256 = 30720$ components to update, which is a large scale problem. Our method clearly performs better. It has shorter time of convergence and higher minimization capacity. It can also be seen through the same figure that initialisation has a great impact on the result of the gradient algorithm.

Our objective was to reduce the coherence between both matrices $\mathbf{\Phi}$ and $\mathbf{\Psi}$ so as to enhance signal reconstruction. Let us examine the incoherence measure achieved by both algorithms, for different initialisation matrices $\mathbf{\Phi}_0$. Table 1 and Table 2 display incoherence values as defined in (2) and (3). Initial values show the incoherence measured before optimization of $\mathbf{\Phi}$. Our sparsity matrix has clearly high mutual coherence with all sorts of initial measurement matrices. Note that our algorithm performs better than the 2-stage gradient method in most cases especially when considering μ_{avg}. Note also that our approach is particularly efficient when dealing with $\mathbf{\Phi}_0$ drawn according to $\mathrm{U}_{[0,1]}$.

Now, we propose to take the $\mathbf{\Phi}$ returned by each algorithm, with different initialisations, to perform reconstruction of 300 test signals belonging to the same manifold of the learning set. Let us first examine the distribution of the off-diagonal elements of the Gram matrix. Fig.3 shows the dis-

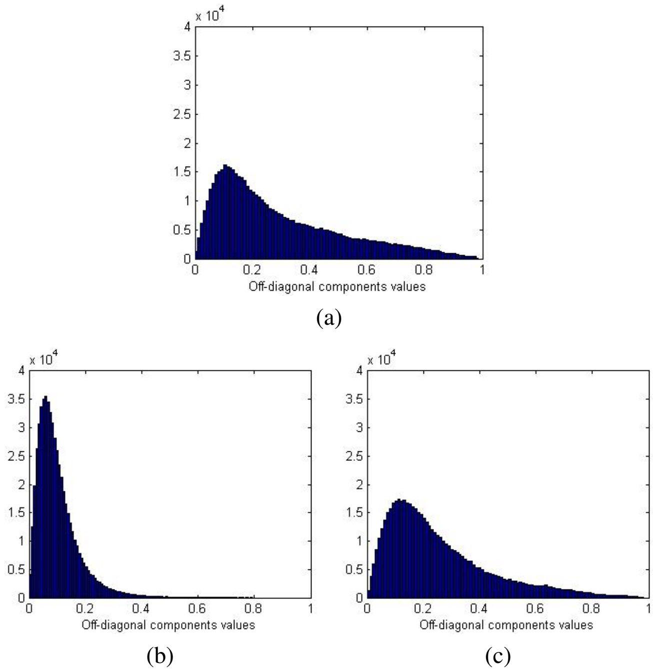

Fig. 3. The histogram of off-diagonal elements of the gram matrix for (a)the initial $\mathbf{\Phi}_0 \sim \mathrm{U}_{[0,1]}$, (b)$\mathbf{\Phi}$ returned by our algorithm and (c)$\mathbf{\Phi}$ returned by 2-step gradient algorithm. $\mathbf{\Psi}$ is the same in all cases (AEMD matrix)

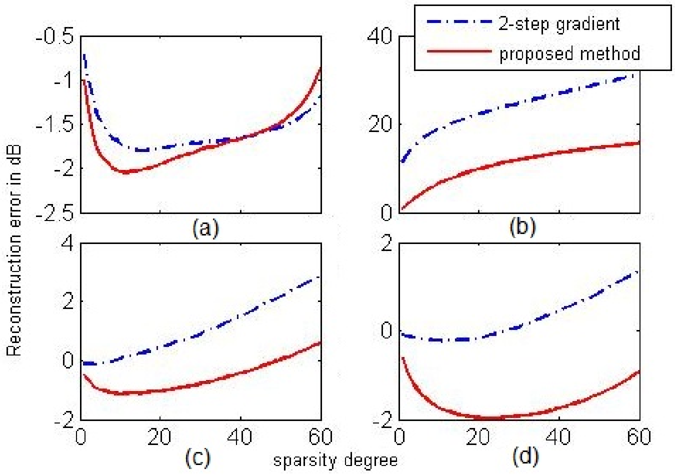

Fig. 4. Reconstruction error averaged over 300 test signals. In each experiment, the $\mathbf{\Phi}$ used is learned from a different initialisation. (a)$\mathbf{\Phi}_0 = I_{M\times N}$, (b)$\mathbf{\Phi}_0 \sim \mathrm{U}_{[0,1]}$, (c)$\mathbf{\Phi}_0 \sim \mathcal{N}(0, \frac{1}{M})$, (d)$\mathbf{\Phi}_0 = Al_1$

$\mathbf{\Phi}_0$	$I_{M\times N}$	$U_{[0,1]}$	$\mathcal{N}(0, \frac{1}{M})$	Al_1
Initial values	1	1	1	1
Proposed approach	0.81	0.80	0.65	0.64
2-stage gradient	0.81	0.98	0.65	0.63

Table 2. μ_{max} values returned by both algorithms for different initialisations.

tribution of $\mathbf{\Phi}$ components obtained from $\mathbf{\Phi}_0 \sim U_{[0,1]}$. As it can be seen, our method makes the distribution closer to 0.
$\widehat{\boldsymbol{x}}$ being the reconstructed signal, we define the reconstruction error as $\varepsilon = \frac{\|\boldsymbol{x}-\widehat{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|}$. Fig.4 shows the reconstruction error achieved with OMP algorithm [12], using different $\mathbf{\Phi}_0$ while varying sparsity degree. The reconstruction is made over 300 test signals. In Fig.4 we display reconstruction error average value over the test set. In all cases, the proposed method achieves reduced reconstruction error compared to the 2-stage gradient method. Note that, the best reconstruction is achieved when $\mathbf{\Phi}_0$ is the identity matrix, and that the worst one iis obtained when $\mathbf{\Phi}_0$ contain uniformly distributed entries. This is true for both methods. Note also that a sparsity $P < 20$ enhances reconstruction in all cases. The best result is got with $\mathbf{\Phi}_0 = I_{M\times N}$ and $P = 10$.

5. CONCLUSION

This paper deals with compressive sensing paradigm and is focusing on designing sparsity and measurement matrices. We first start by learning the sparsity matrix using Hilbert transformation and empirical mode decomposition. The learned matrix is therefore composed of complex components and has a good sparsifying property. Having the sparsity matrix learnt, we use a gradient method to learn an adaptive measurement matrix. The goal is to reduce the mutual coherence of both matrices. Our method integrates the dictionary normalization step into the function to optimize. Simulations prove that it supports large scale problems, has good convergence properties and achieves good performance with respect to incoherence. Experiments on environmental sounds show that the whole approach outperforms previous method in signal reconstruction. Further, we will be interested in applying this approach in a classification oriented measurement matrix design.

REFERENCES

[1] S. Kunis and H. Rauhut, “Random sampling of sparse trigonometric plynomials, ii.orthogonal matching pursuit versus basis pursuit,” *Foundations of Computational Mathematics*, vol. 8, pp. 737–763, 2008.

[2] K. Engan, S.O. Aase, and J. Hakon Husoy, “Method of optimal directions for frame design,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1999, vol. 5, pp. 2443–2446.

[3] M. Aharon, M. Elad, and A. Bruckstein, “Svdd: an algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE Transactions on Signal Processing*, vol. 54, pp. 4311–4322, 2006.

[4] M.G. Jafari and M.D. Plumbley, “Fast dictionary learning for sparse representations of speech signals,” in *Journal of Selelected Topics in Signal Processing*, 2011, pp. 1025–1031.

[5] B. Bouchhima, R. Amara, and M. Turki-Hadj Alouane, “Dictionary learning using emd and hilbert transform for sparse modeling of environmental sounds,” in *Proceedings of the International Conference on Artificial Intelligence and Pattern Recognition*, 2014, pp. 104–110.

[6] M. Elad, “Optimized projections for compressed sensing,” *IEEE Transaction on Signal Processing*, vol. 55, pp. 5695 – 5702, 2007.

[7] J.M. Duarte-Carvajalino and G. Sapiro, “Learning to sense sparse signals: simultaneous sensing matrix and sparsifying dictionary optimization,” *IEEE Transaction on Image Processing*, vol. 18, pp. 1395 – 1408, 2009.

[8] E. Osherovich, “Designing incoherent dictionaries for compressed sensing: algorithm comparison,” *CoRR*, vol. abs/1206.4192, 2012.

[9] J. Xu, Y. Pi., and Z. Cao, “Optimized projection matrix for compressive sensing,” *EURASIP Journal on Advances in Signal Processing*, 2010.

[10] V. Abolghasemi, S. Ferdowsi, B. Makkiabadi, and S. Sanei, “On optimization of the measurement matrix for compressive sensing,” in *European Signal Processing Conference*, 2010, pp. 427–431.

[11] N. E.Norden Huang and Samuel.S Shen, *Hilbert-Huang transform and its applications*, vol. 5, World Scientific Publishing Co. Pte. Ltd, 2008.

[12] G. Davis, S. Mallat, and Z. Zhang, “Adaptive time-frequency decompositions with matching pursuits,” *Optical Engineering*, vol. 33, 1994.