BLIND SOURCE SEPARATION OF MEDIAL TEMPORAL DISCHARGES VIA PARTIAL DICTIONARY LEARNING

Shahrzad Shapoori, Saeid Sanei, and Wenwu Wang

Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, UK
{s.shapoori, w.wang & s.sanei}@surrey.ac.uk

ABSTRACT

Sparsity is known to be very beneficial in blind source separation (BSS). Even if data is not sparse in its current domain, it can be modelled as sparse linear combinations of atoms of a chosen dictionary. The choice of dictionary that sparsifies the data is very important. In this paper the dictionary is partly pre-specified based on chirplet modelling of various kinds of real epileptic discharges, and partly learned using a dictionary learning algorithm. The dictionary which includes a fixed and a variable (i.e. learned) part, is incorporated into a source separation framework to extract the closest source to the source of interest from the mixtures. Experiments on synthetic mixtures of real data consisting of epileptic discharges are used to evaluate the proposed method, and the results are compared with a traditional BSS algorithm.

Index Terms— partial dictionary learning, sparsity, BSS, medial temporal discharges, epilepsy

1. INTRODUCTION

Prediction of seizure is very beneficial in diagnosis of many neurological disorders, such as epilepsy. Long before the occurrence of seizure there are small medial temporal discharges in the intracranial Electroencephalogram (EEG). The aforementioned discharges, also known as epileptic spikes, look much more distinct in intracranial than in scalp EEG recordings. Scalp EEG can be considered as a mixture of intracranial sources, nonlinearity of head, and noise of recording devices and environment [1].

To estimate intracranial EEG, blind source separation (BSS) technique can be used, by treating scalp EEG as a set of mixtures and intracranial EEG as original sources, with no or little prior knowledge about the sources, as follows:

$$\mathbf{Y} = \mathbf{AX} + \mathbf{V} \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^{m\times N}$ is the observation matrix, $\mathbf{X} \in \mathbb{R}^{n\times N}$ is the source matrix, and $\mathbf{A}$ is the $m \times n$ mixing matrix. The additive noise caused by the instrumental noise or imperfection of the model which is denoted by $\mathbf{V}$ is of size $m \times N$. The aim of BSS is to estimate both $\mathbf{X}$ and $\mathbf{A}$ from $\mathbf{Y}$.

There is no exact solution to the BSS problem. Incorporating constraints into the separation process has shown to be useful in obtaining the desired solutions. Independent component analysis (ICA) [2], uses statistical independency and non-gaussianity, and the minimization of mutual information as constraints.

Sparsity is another constraint, which has been used recently for source separation by many researchers [3], [4], [5], in particular for the underdetermined case, where the number of sources is greater than the number of mixtures. Apart from the signal being sparse in its current domain, it can also be sparsified using a dictionary, in the sense that only a few atoms of the dictionary are chosen to represent the signal. This dictionary can be composed of different types of signals. It can be fixed or can be learnt from the mixtures themselves.

The way the signals are represented have direct relevance to the choice of dictionary. Dictionary atoms, which are a set of basis signals, are used to decompose the data. Linear combinations of these atoms (elements) are used to uniquely represent different signals. It is, however, a practical challenge to choose suitable dictionaries for different application problems.

There are mainly two types of dictionaries that have been considered in the literature, namely, pre-defined (or pre-specified) dictionaries with a mathematical transform such as the traditional dictionaries obtained by wavelet and Fourier transforms, and more recently the dictioanries that are learned from training data using a machine learning algorithm. The pre-defined dictionaries are often easy to implement and computationally efficient, while the learned dictionaries, on the other hand, may offer better fit to the signals to be represented despite of its relatively higher computational complexity [6].

In this paper, we propose a method for source separation where the dictionary is partly fixed and partly learned. The fixed part is composed of chirplet reconstructed epileptic discharges, and the learned part is adapted from the source signal itself. The main motivation behind our idea is that the epileptic discharges often have some basic structures which can be

This work has been supported by the EPSRC, UK, Grant No. EP/K005510/1.

978-1-4673-7454-5/15/$31.00 ©2015 IEEE

well defined with templates as in the fixed dictionary, and at the same time, partially learning the dictionary will allow the variations in the epileptic discharges to be captured by the atoms adapted from training data. Such a mixed dictionary has the potential to offer better performance for source separation as demonstrated in our work where a joint dictionary learning and source separation framework is employed for the extraction of the source of interest. Some preliminary results are provided to demonstrate the effectiveness of the proposed method based on synthetic mixtures of real EEG data, as compared with a popular and traditional BSS method FastICA.

2. PROBLEM DEFINITION

2.1. Creation of Fixed Dictionary

The fixed part of the dictionary is pre-specified based on different shapes of real epileptic spikes scored by clinician experts in the intracranial EEG. These are reconstructed spikes modelled using a restricted number of chirplets. Some of these reconstructed signals are shown in Fig. 1.

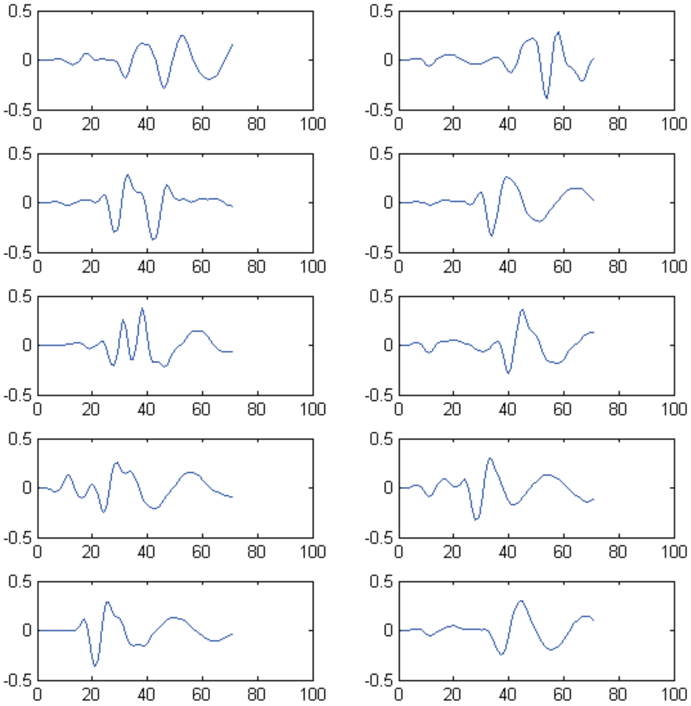

Fig. 1. Some of the dictionary components, based on chirplet modellings of selected pre-ictal discharges in intracranial EEG.

As shown in Fig. 1, epileptic spikes are usually composed of a sharp peak followed by a slow wave [7].

2.2. Sparse Recovery

Calculating the coefficients of the sparse matrices in order to represent the signal using the dictionary is called sparse recovery or atom decomposition. This is often achieved by solving one of the following optimisation problems. This is usually done by a matching pursuit algorithm [8].

$$\hat{\mathbf{s}} = \min_{\mathbf{s}} \|\mathbf{s}\|_0 \;\; subject\ to \;\; \mathbf{x} = \mathbf{D}\mathbf{s} \tag{2}$$

$$\hat{\mathbf{s}} = \min_{\mathbf{s}} \|\mathbf{s}\|_0 \;\; subject\ to \;\; \|\mathbf{x} - \mathbf{D}\mathbf{s}\|_2 \leq \epsilon \tag{3}$$

where $\|.\|_0$ is the l_0-norm and counts the number of non-zeros in its argument, **s** is the vector of sparse coefficients, **x** is the signal to be decomposed, and **D** is the dictionary. The basic idea is that every signal can be represented as a linear combination of a few number of dictionary atoms (columns). Model (3) is an alternative version of model (2) taking into account modelling errors and noise.

Because exactly determining the sparsest representation of the signal is an NP-hard problem [9], approximate solutions are considered instead of precise ones. Matching pursuit (MP) [8] and orthogonal matching pursuit (OMP) [10] algorithms select the dictionary atoms consecutively. Computational efficiency and easiness of implementation are two of the major advantages of these algorithms.

These methods involve calculating the inner product between the signal and the dictionary atoms. Some least square solvers are probably used as well. Firstly the contribution due to the atom with largest inner product with the signal is subtracted from the signal. Then the contribution due o the atom with second largest inner product with the signal is subtracted, and the procedure continues until the signal is completely decomposed. Both (2) and (3) are addressed by changing the algorithm stopping rule. The process of updating all the extracted coefficients after each step, is the main thing that differentiates MP from OMP. This step is performed in OMP by calculating the orthogonal projection of the signal onto the set of atoms which are selected after each iteration. OMP requires more calculation. However, it leads to better results than MP. These algorithms and similar methods are able to recover the solution, if it is sparse enough.

2.3. Source Separation

If in the BSS model in section I Eq. (1), the sources of interest and the mixtures are considered intracranial epileptic discharges and scalp recordings of epileptic patients respectively, in vector form the model can be re-written as:

$$\underline{\mathbf{y}} = (\mathbf{I} \otimes \mathbf{A})\underline{\mathbf{x}} + \underline{\mathbf{v}} \tag{4}$$

In the above equation, $\underline{\mathbf{x}} = vec(\mathbf{X})$ and $\underline{\mathbf{y}} = vec(\mathbf{Y})$ are column vectors of length *nN* and *mN* respectively, in which *n* is the number of scalp channels, *m* is the number of sources, and

N is the number of samples. $(\mathbf{I}\otimes\mathbf{A})$ of size $mN\times nN$ is a block diagonal matrix, and $\otimes$ is the Kronecker product symbol.

3. THE PROPOSED ALGORITHM

3.1. The Overall Cost Function

The overall source separation problem via partial dictionary learning is expressed as:

$$\min_{\{s_i\},D,\underline{x},A} \lambda\|\underline{\mathbf{y}} - (\mathbf{I}\otimes\mathbf{A})\underline{\mathbf{x}}\|_2^2 + \sum_{i=1}^{p}[\mu_i\|\mathbf{s}_i\|_0 + \|\mathbf{D}\mathbf{s}_i - \Re_i\underline{\mathbf{x}}\|_2^2] \tag{5}$$

where λ and μ control the noise power and sparsity degree, respectively. $\mathbf{D}$ is the dictionary of size $r\times k$ which contains normalised columns, also called atoms. $\{\mathbf{s}_i\}$ are sparse coefficients of length k. For simplicity the source vector $\mathbf{x}$ is divided into patches of length r, and the i-th patch from $\mathbf{x}$ is shown by vector $\Re_i\mathbf{x}$. Furthermore, the total number of patches is shown by p. $\mathbf{A}$ shows the mixing matrix. The constrained problem mentioned in section II, has changed to an unconstrained problem by the use of penalty terms.

3.2. Source Separation via Partial Sparsity-Based Dictionary Learning

The cost function (5) is minimized using an alternating method by keeping all but one of the unknowns fixed at a time. Before the learning process, the mixing matrix $\mathbf{A}$ is initialized by a random matrix of suitable size, the source vector $\mathbf{x}$ is initialized by $\mathbf{x} = \mathbf{A}^T\mathbf{y}$, and the variable part of the dictionary $\mathbf{D}$ is initialised by cosine waves. Segments of $\mathbf{x}$ are handled one at a time. The OMP algorithm estimates the sparse coefficients $\{\mathbf{s}_i\}_{i=1}^{p}$ for each segment $\Re_i\mathbf{x}$, gathering one atom at a time, and stopping when the error goes below a threshold. The values for μ_i are chosen implicitly. In other words, in this stage a sliding window applies sparse coding on each segment of data one at a time. Given all $\mathbf{s}_i$ values, $\mathbf{A}$, and $\mathbf{x}$, the dictionary $\mathbf{D}$ can now be updated using a sequence of K-SVD operations as in [11]. The dictionary is updated columnwise using singular value decomposition (SVD). Once this is done, the signal $\mathbf{x}$ is updated by keeping all $\mathbf{s}_i$ values, $\mathbf{A}$, and $\mathbf{D}$ fixed. In order to obtain $\mathbf{x}$, the gradient of (5) is calculated and set to zero:

$$0 = \lambda(\mathbf{I}\otimes\mathbf{A})^T((\mathbf{I}\otimes\mathbf{A})\underline{\mathbf{x}} - \underline{\mathbf{y}}) + \sum_{i=1}^{p}\Re_i^T(\Re_i\underline{\mathbf{x}} - \mathbf{D}\mathbf{s}_i) \tag{6}$$

which becomes:

$$\hat{\underline{\mathbf{x}}} = (\lambda(\mathbf{I}\otimes\mathbf{A})^T(\mathbf{I}\otimes\mathbf{A}) + \sum_{i=1}^{p}\Re_i^T\Re_i)^{-1}\cdots \cdots(\lambda(\mathbf{I}\otimes\mathbf{A})^T\underline{\mathbf{y}} + \sum_{i=1}^{p}\Re_i^T\mathbf{D}\mathbf{s}_i). \tag{7}$$

Solving equation (7) leads to the solution of $\mathbf{x}$. In order to estimate the mixing matrix $\mathbf{A}$, all unknowns except $\mathbf{A}$ are considered fixed. As shown below, the mixtures signal $\mathbf{y}$ and the source signal $\mathbf{x}$, which is computed in the previous step, are used. Simplifying (5) by changing the first quadratic term into normal matrix product gives:

$$\min_{\mathbf{A}} \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_F^2. \tag{8}$$

This minimization problem is solved simply by taking the pseudo inverse of $\mathbf{X}$ and then estimate $\mathbf{A}$ as:

$$\hat{\mathbf{A}} = \mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1} \tag{9}$$

Now, the output source $\mathbf{x}$ is computed. However, an update of the output source also updates the sparse coefficients, $\{\mathbf{s}_i\}_{i=1}^{p}$. Therefore, the steps of estimating $\{\mathbf{s}_i\}$, $\mathbf{D}$, $\mathbf{x}$, and $\mathbf{A}$ have to be alternatively repeated for a suitable number of iterations, in order to minimize (5) and obtain a source close to the desired solution.

4. EXPERIMENTAL RESULTS

4.1. Data

Synthetic mixtures of real data consisting of epileptic discharges, are used in our experiments. It is produced by linear mixing of a piece of intracranial signal, which contains epileptic discharges, and a random signal, together with Gaussian noise. This makes the overall mixing system underdetermined.

4.2. Results and Discussion

As depicted, the signal in Fig. 2(a), which shows an intracranial discharge is linearly mixed with the signal in Fig. 2(b), and different noise levels are added to them in order to create the mixtures, one set of them is shown in Fig. 2(c) and 2(d). The amplitude of intracranial discharge is made 10 times less than the amplitude of the random signal, before mixing. This has been done in order to make the simulated data more similar to the real scalp signals. In reality, epileptic discharges in the intracranial signals have much smaller amplitudes than the scalp EEG.

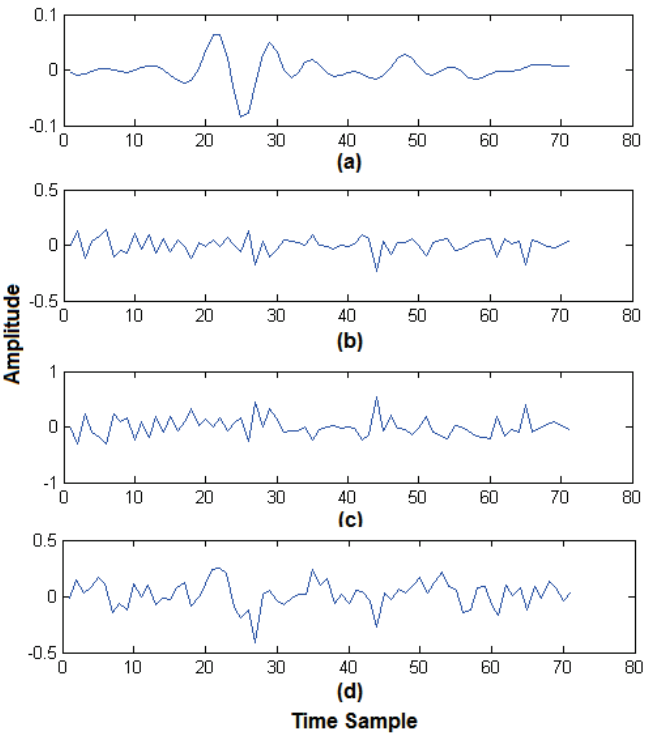

Fig. 2. Simulated original sources and mixtures; (a) Original epileptic discharge (b) Random signal (c) Mixture 1 (d Mixture 2.

Conventional BSS, in this case FastICA, has been applied to these mixtures and the result is shown in Fig. 3.

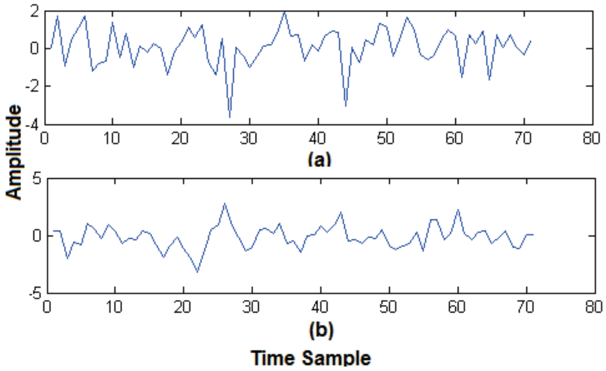

Fig. 3. Separated sources using FastICA; (a) Source 1 (b) Source 2.

The proposed method also has been applied to these mixtures and the output source result is shown in Fig. 4. It is shown that the original source has been extracted from the set of mixtures with just a little noise, and the result is better than that of the FastICA.

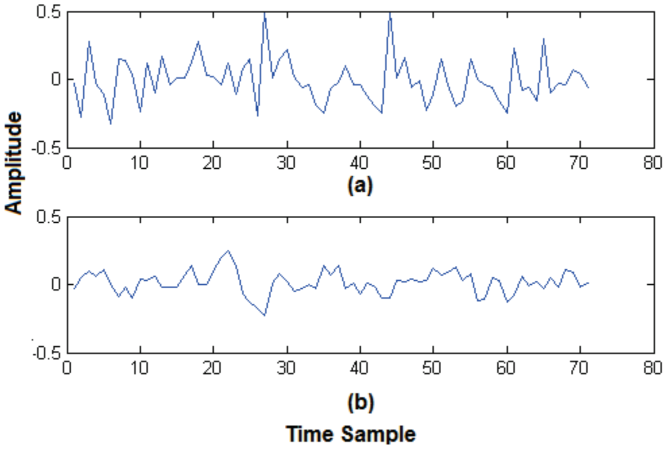

Fig. 4. Proposed method output sources; (a) Source 1 (b) Source 2.

Also, a comparison of the root mean square error (RMSE) between the desired source and the extracted source by FastICA and the proposed method is shown in Fig. 5.

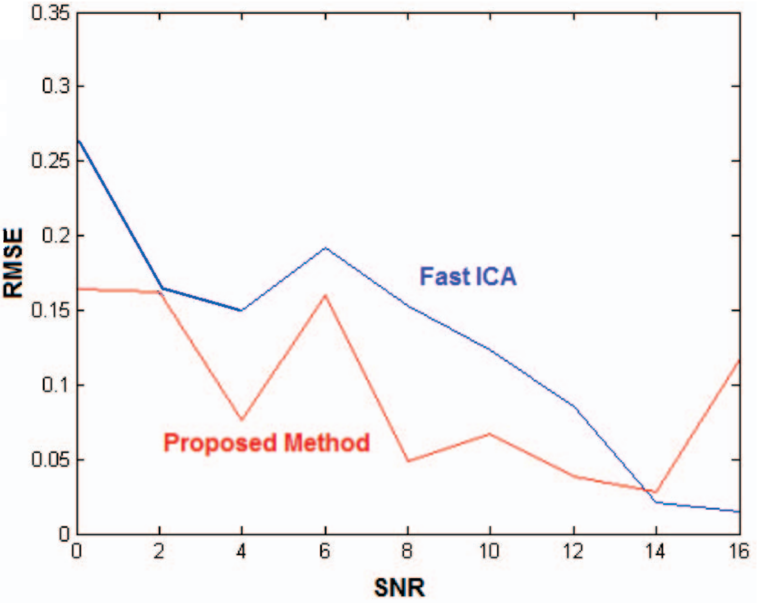

Fig. 5. Comparison of FastICA and the proposed method with different SNRs.

5. CONCLUSIONS

In this paper, a source separation method based on partial dictionary learning has been proposed. The dictionary is partly pre-specified with different epileptic discharges, which are produced by chirplet modelling of the actual spikes, and is partly learned through the dictionary learning algorithm. Iteratively estimating the sparse coefficients, the dictionary, and the mixing matrix, the closest source to the original source from the mixtures is extracted. The method has been tested on synthetic mixtures of real data which consists of epileptic discharges, and the results are compared with FastICA which is a traditional BSS method. In future work, the algorithm will be evaluated on real EEG data.

6. REFERENCES

[1] S. Sanei, *Adaptive Processing of Brain Signals*. John Wiley & Sons, 2013.

[2] A. Hyvärinen, J. Hurri, and P. O. Hoyer, "Independent component analysis," in *Natural Image Statistics*. Springer, 2009, pp. 151–175.

[3] A. M. Bronstein, M. M. Bronstein, M. Zibulevsky, and Y. Y. Zeevi, "Sparse ica for blind separation of transmitted and reflected images," *International Journal of Imaging Systems and Technology*, vol. 15, no. 1, pp. 84–91, 2005.

[4] V. Abolghasemi, S. Ferdowsi, and S. Sanei, "Blind separation of image sources via adaptive dictionary learning," *IEEE Transactions on Image Processing*, vol. 21, no. 6, pp. 2921–2930, 2012.

[5] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: from coding to source separation," *IEEE Proceedings*, vol. 98, no. 6, pp. 995–1005, 2010.

[6] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *IEEE Proceedings*, vol. 98, no. 6, pp. 1045–1057, 2010.

[7] N. Kissani, G. Alarcon, M. Dad, C. Binnie, and C. Polkey, "Sensitivity of recordings at sphenoidal electrode site for detecting seizure onset: evidence from scalp, superficial and deep foramen ovale recordings," *Clinical Neurophysiology*, vol. 112, no. 2, pp. 232–240, 2001.

[8] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[9] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *Constructive Approximation*, vol. 13, no. 1, pp. 57–98, 1997.

[10] Y. C. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *IEEE Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*. IEEE, 1993, pp. 40–44.

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