COUPLED DICTIONARY LEARNING FOR MULTIMODAL DATA: AN APPLICATION TO CONCURRENT INTRACRANIAL AND SCALP EEG

Loukianos Spyrou and Saeid Sanei

Department of Computer Science, University of Surrey, U.K.

ABSTRACT

This paper focuses on learning a coupled dictionary between multimodal datasets where the data of different modes can be described as a function of each other. Our method is able to reconstruct the data of one mode by using the data of another mode. This provides the advantage on applications that low-quality data are generally available and high-quality data are not. We employ a concurrent intracranial and scalp EEG dataset, to learn a dictionary and a mapping function between the two modalities. The aim is to infer the intracranial from only the scalp EEG by using that dictionary and mapping function. The novelty of this work is the development of an algorithm that obtains an optimal coupled dictionary, sparse coefficients and the mapping function between modalities.

Index Terms— coupled dictionary learning, sparsity, intracranial, EEG, superresolution

1. INTRODUCTION

Detection of interictal epileptiform discharges (IED) from scalp EEG (sEEG) is highly desirable for a variety of clinical and research fields. Since intracranial EEG (iEEG) IEDs correspond more closely to the true brain activity their accurate estimation enables better diagnosis and accurate estimation of brain functions. IEDs present on sEEG are considered degraded by interference, attenuation, blurring, or delay as compared to those on iEEG [1].

Dictionary learning is a widely used methodology in signal processing and machine learning research. It involves modelling data as a linear combination of basis elements called atoms. The main advantage of such a method is that if the dictionary describes the target signal accurately, noise in the reconstructed signal will be reduced since the noise does not fit the dictionary. Sparse coding is one of the methodologies that the atoms are combined and has numerous applications in fields where data vectors are desired to consist only of a small number of atoms. For a review of dictionary learning and sparse coding please refer to [2]. Dictionary learning methods have also been applied in the EEG field [3] and for processing of epileptic data [4].

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

The main novelty of this work is that we learn a common dictionary and its mapping function between coupled modalities. We assume that the same sparse approximation is valid for both modes and that the signal from one can be reconstructed by estimating the sparse coefficients from the other. The advantage is that when only the data from one modality are available and their sparse representation is computed, we can reconstruct the data from the other modality. There are many real-world cases where high-quality data are to be reconstructed from low-quality data.

There have been only a few studies with coupled dictionaries in the literature. In [5], [6] and [7] different dictionaries are learned but a common mapping between the sparse coefficients is estimated. In a superresolution context [8], the coupled dictionaries share the same coefficients but are allowed to take any form and no mapping between them is estimated. In [9], the dictionaries are jointly learned which describe different aspects of the same image. In [10], the dictionaries in temporal and DFT domains are learned jointly together with a mapping between the sparse coefficients. Our contributions can be summarised as follows:

- estimation of a linked dictionary for both modalities
- estimation of a linear mapping between the dictionaries that enables the reconstruction of one from the other by using shared sparse approximation coefficients

In Section 2.1 we describe the theory behind dictionary learning and sparse coding. Section 2.2 derives our coupled dictionary learning approach while in Section 2.3 we describe the performance of our method and the comparison procedures we follow. In Section 3 we show results for both simulated and real epileptic data. Section 4 concludes the paper.

2. METHODS

2.1. Dictionary Learning and Sparse Coding

Dictionary learning involves the process of estimating a dictionary $\mathbf{D} \in \mathbb{R}^{N \times m}$ that can accurately describe a signal $\mathbf{y} \in \mathbb{R}^N$ by estimating m atoms each denoted by $\mathbf{d}_j$. The signal is expressed as a linear combination, $\mathbf{a}$, of a small number atoms where if $m > N$ the dictionary is overcomplete and

spans the signal subspace. The signal is expressed as:

$$\mathbf{y} = \mathbf{D}\mathbf{a} = \sum_{j=1:m} \mathbf{d}_j \mathbf{a}_j \tag{1}$$

where typically it is desired that the coefficients are sparse so that the reconstructed signal contains only a few significant atoms. The joint dictionary learning and sparse approximation is therefore formulated as:

$$\underset{\mathbf{D},\mathbf{a}}{\operatorname{argmin}} ||\mathbf{y} - \mathbf{D}\mathbf{a}|| \;\; s.t. \;\; ||\mathbf{a}||_{\mathbf{0}} \tag{2}$$

where $||.||_p$ denotes the l_p norm which for the case of the l_0-norm is NP hard and nonconvex. Approximate algorithms are traditionally applied and in this work we use nonnegative least squares methods and Lagrange multipliers to obtain the sparse coefficients and the dictionary. The Sparse Representation Toolbox was used and modified for all computations [11].

2.2. Coupled Dictionary Learning

In this work we provide a new coupled dictionary learning with sparse approximation (CDLSA) formulation that takes advantage of datasets, where coupled signals from different modalities are linked by a linear operator $\mathbf{C}$. Examples of such signals can be low and high quality images of the same object, measurements from different devices and in our case concurrent intracranial and scalp EEGs. Let $\mathbf{x}_i \in \mathbb{R}^N$ be the high quality signal and $\mathbf{x}_s \in \mathbb{R}^M$ be the low quality signal with:

$$\mathbf{x}_s = \mathbf{C}\mathbf{x}_i + \mathbf{W} \tag{3}$$

where $\mathbf{C} \in \mathbb{R}^{N \times M}$ and $\mathbf{W}$ denotes measurement and modelling noise. We wish to estimate a dictionary $\mathbf{D} \in \mathbb{R}^{N \times m}$ that is common between the modalities in a similar fashion:

$$\mathbf{D}_s = \mathbf{C}\mathbf{D}_i \equiv \mathbf{C}\mathbf{D} \tag{4}$$

The usual sparse dictionary learning can converted to our problem as such:

$$\mathbf{J}(\mathbf{D}, \mathbf{a}) = ||\mathbf{x}_s - \mathbf{C}\mathbf{D}\mathbf{a}||_{\mathbf{2}}^{\mathbf{2}} + ||\mathbf{x}_i - \mathbf{D}\mathbf{a}||_{\mathbf{2}}^{\mathbf{2}} + \lambda||\mathbf{a}||_{\mathbf{1}} \tag{5}$$

Both modalities use the same sparse coefficients since the same processes generate both modalities' signals. We convert Eq. (5) to the standard form as:

$$\mathbf{J}(\mathbf{D}, \mathbf{a}) = ||\hat{\mathbf{x}} - \hat{\mathbf{D}}\mathbf{a}||_{\mathbf{2}}^{\mathbf{2}} + \lambda||\mathbf{a}||_{\mathbf{1}} \tag{6}$$

where $\hat{\mathbf{x}} = \begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_i \end{bmatrix}$ and

$$\hat{\mathbf{D}} = \begin{bmatrix} \mathbf{C} \\ \mathbf{I} \end{bmatrix} \mathbf{D} = \hat{\mathbf{C}}\mathbf{D} \tag{7}$$

Since we are using the standard form, any dictionary learning and sparse decomposition method can be used to solve Eq. (6). The coupled dictionary $\mathbf{D}$ admits an analytic expression by considering L examples from our linked datasets $\mathbf{X_i} \in \mathbb{R}^{N \times L}$, $\mathbf{X_s} \in \mathbb{R}^{M \times L}$ and $\hat{\mathbf{X}} \in \mathbb{R}^{(N+M) \times L}$. Eq. (6) can be expanded as follows:

$$\mathbf{J}(\mathbf{D}, \mathbf{a}) = \mathbf{X}^{\mathbf{T}}\hat{\mathbf{X}} + \mathbf{A}^{\mathbf{T}}\mathbf{D}^{\mathbf{T}}\hat{\mathbf{C}}^{\mathbf{T}}\hat{\mathbf{C}}\mathbf{D}\mathbf{A} - \mathbf{A}^{\mathbf{T}}\mathbf{D}^{\mathbf{T}}\mathbf{C}^{\mathbf{T}}\hat{\mathbf{X}} \tag{8}$$

where $\mathbf{A}$ contains the sparse coefficients for each example. By taking the derivative *w.r.t.* $\mathbf{D}$ and setting that to zero, we obtain the solution for the dictionary $\mathbf{D}$:

$$\frac{\partial \mathbf{J}(\mathbf{D}, \mathbf{a})}{\partial \mathbf{D}} = (\hat{\mathbf{C}}^{\mathbf{T}}\hat{\mathbf{C}})\mathbf{D}\mathbf{A}\mathbf{A}^{\mathbf{T}} - \hat{\mathbf{C}}^{\mathbf{T}}\hat{\mathbf{X}}\mathbf{A}^{\mathbf{T}} = \mathbf{0} \tag{9}$$

$$\mathbf{D} = (\hat{\mathbf{C}}^{\mathbf{T}}\hat{\mathbf{C}})^{-\mathbf{1}}\hat{\mathbf{X}}\mathbf{A}^{\mathbf{T}}(\mathbf{A}\mathbf{A}^{\mathbf{T}})^{-\mathbf{1}} \tag{10}$$

The step for learning the sparse coefficients is performed for each example sequentially by converting Eq. (6) to the format $\mathbf{a}^{\mathbf{T}}\mathbf{H}\mathbf{a} + \mathbf{a}^{\mathbf{T}}\mathbf{g}$ $\mathbf{s.t.}$ $\mathbf{a} \geq \mathbf{0}$. For each $\mathbf{a} \in \mathbf{A}$ we solve the standard nonneagive quadratic programm (NNQP):

$$\mathbf{a}^{\mathbf{T}}(\hat{\mathbf{D}}^{\mathbf{T}}\hat{\mathbf{D}})\mathbf{a} - \mathbf{a}^{\mathbf{T}}(\hat{\mathbf{D}}^{\mathbf{T}}\hat{\mathbf{x}} - \lambda) \;\; s.t. \; \mathbf{a} \geq \mathbf{0} \tag{11}$$

where λ controls the sparsity of the solution. The algorithm alternates between estimating $\mathbf{D}$ and $\mathbf{A}$ until convergence determined by the residual in Eq. (6). The linear function $\mathbf{C}$ can also be estimated by $\operatorname{argmin}_C ||\mathbf{X_s} - \mathbf{C}\mathbf{D}\mathbf{A}||$ which is solved approximately for:

$$\mathbf{C} = \mathbf{X_s}(\mathbf{D}\mathbf{A})^{\dagger} \tag{12}$$

where † denotes the pseudoinverse. When the optimisation of $\mathbf{C}$ is performed the algorithms alternates between estimating $\mathbf{D}$, $\mathbf{A}$ and $\mathbf{C}$.

2.3. Performance Evaluation

Since the overarching aim is to reconstruct the high quality signals $\mathbf{x}_i$ from low quality signals $\mathbf{x}_s$ when e.g. $\mathbf{x}_i$ are not available but $\mathbf{x}_s$ are, we perform sparse coding with the dictionary $\mathbf{D}_s$ corresponding to the low quality signal only. Recall Eq. (4) where $\mathbf{D}_s = \mathbf{C}\mathbf{D}$ with $\mathbf{D}$ and $\mathbf{C}$ given by the coupled dictionary algorithm. The sparse coefficients for a test signal $\mathbf{x}_s^{tst}$ are given by:

$$\underset{\mathbf{a}^{tst}}{\operatorname{argmin}} ||\mathbf{x}_s^{tst} - \mathbf{C}\mathbf{D}\mathbf{a}^{tst}||_{\mathbf{2}}^{\mathbf{2}} + \lambda||\mathbf{a}^{tst}||_{\mathbf{1}} \tag{13}$$

Then, the same $\mathbf{a}^{tst}$ is used to reconstruct $\mathbf{x}_{\mathbf{i}}^{tst}$ by dropping $\mathbf{C}$:

$$\mathbf{y}_i^{tst} = \mathbf{D}\mathbf{a}^{tst} \tag{14}$$

The first motivation behind our proposed method is that the reconstructed test signal of Eq. (14) can be a better approximation to the true high quality signal than a signal based on a dictionary learned directly from $\mathbf{x}_i$ with a subsequent estimation of the linear function $\mathbf{C}'$. In other words we compare the signal obtained by our method with a signal:

$$\mathbf{z}_i^{tst} = \mathbf{D}_i'\mathbf{q}^{tst} \tag{15}$$

where $\mathbf{D}_i'$ has been obtained by traditional dictionary learning for only the high quality signal $\mathbf{x}_i^{trn}$ and $\mathbf{q}^{tst}$ are the coefficients of the test set. The mapping function is obtained in a similar way to the CDLSA method:

$$\mathbf{C}' = \mathbf{X}_s^{trn}(\mathbf{D}_i\mathbf{Q}_i^{trn})^{\dagger} \tag{16}$$

where $\mathbf{Q}_i^{trn}$ are the sparse coefficients of all trials obtained from the training set. The sparse coefficients of the test signal are obtained as such:

$$\underset{\mathbf{q}^{tst}}{\operatorname{argmin}} ||\mathbf{x}_s^{tst} - \mathbf{C}'\mathbf{D}_i'\mathbf{q}^{tst}||_{\mathbf{2}}^{\mathbf{2}} + \lambda||\mathbf{q}^{tst}||_{\mathbf{1}} \tag{17}$$

Hence, the reconstructed signal $\mathbf{y}_i^{tst}$ from our proposed method can be compared with $\mathbf{z}_i^{tst}$. We denote the method that obtains $\mathbf{z}_i^{tst}$ as DLSA+C.

We also compare our method with that of [8] for which the two dictionaries can take a different form i.e. instead of Eq. (18) the coupled dictionary can take any form and the function $\mathbf{C}$ is not estimated:

$$\hat{\mathbf{D}} = \begin{bmatrix} \mathbf{D}_s \\ \mathbf{D}_i \end{bmatrix} \tag{18}$$

3. RESULTS

3.1. Dataset

The study included the data from 10 patients with scalp EEG recordings and simultaneous intracranial multicontact foramen ovale (FO) electrode bundles in the Department of Clinical Neurophysiology at Kings College Hospital [12]. Two flexible bundles of 6 electrodes each were inserted through the left and right FO [13]. Cable telemetry of 32 channels was used for data acquisition. Data were digitised at 200 Hz and bandpass filtered in the device ([0.3 70]Hz). From each patient, a period of 20 min of intracranial EEG recordings were transcribed onto a digital file. A clinician visually inspected the iEEG data and marked the timepoints of epileptic spikes. Both iEEG and sEEG were sliced (i.e. trials) according to those timepoints in a $\pm 162.5ms$ window and further used for analysis. The data were pre-processed by further filtering at the [1 45]Hz range and common average referenced separately for iEEG and sEEG. Baseline drifts were removed by performing first order linear detrending. Each multichannel trial of both modalities was subsequently vectorised. This step was performed such that iEEG and sEEG admit the same representation. We split each patient's dataset in two equal parts, a training set where the dictionary learning was performed and a test set where the reconstructed signals were obtained.

3.2. Epileptic Data

All the methods were applite to each subject separately and the error on each was computed. The error was computed as the average mean square error over each subject's trials on the test set. Note that each subject has a different number of trials ranging from 50 to 900. Initially, we performed a grid search on the training set to obtain reasonable parameters for the number of atoms m. The error on the training set gradually decreased as m increased. However, a good choice for the numbers of atoms was set at $m = 10$ since the test error didn't decrease any further after that. Similarly, the sparsity parameter was set at $\lambda = 0.05$.

The first test we performed was to create a set of semi-simulated data for each subject. The scalp EEG was a simulated as a noisy version of the intracranial:

$$\mathbf{x}_s = \mathbf{C}\mathbf{x}_i + \sigma\mathbf{W} \tag{19}$$

where $\mathbf{C}$ was a random propagation function from the intracranial to the scalp EEG, $\mathbf{W}$ standard gaussian noise and σ controls the variance of the noise. For each subject, we calculated the sparse approximation on the test set of the scalp signals. Subsequently, we calculated the decrease in test error of the reconstructed intracranial signal between our method, the method in [8] as compared with the DLSA+C method described in Section 2.3. For increasing σ we show the group average test error in Figure 1. In this case the function $\mathbf{C}$ was estimated from the data. For known and fixed $\mathbf{C}$, we obtained Figure 2. Initially, as σ increases, both our method and the one in [8] perform better than the benchmark method DLSA+C for both cases. After a point, the improvement is longer effective since noise has degraded the scalp signal substantially.

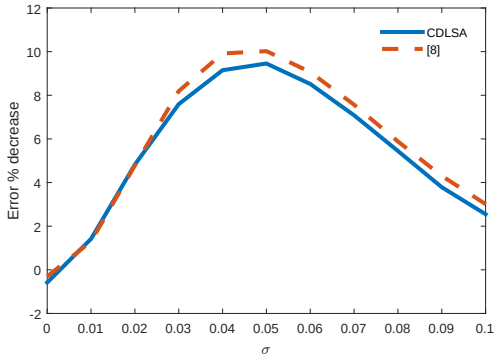

Fig. 1. Group average percentage decrease in error on the test set between our CDLSA method and the method in [8] with the DLSA+C method described in Section 2.3. The mapping $\mathbf{C}$ was estimated from the data.

For $m = 10$, $\lambda = 0.05$ we obtained the results in Table 1 which shows the reconstructed errors of the intracranial signal on the test set.

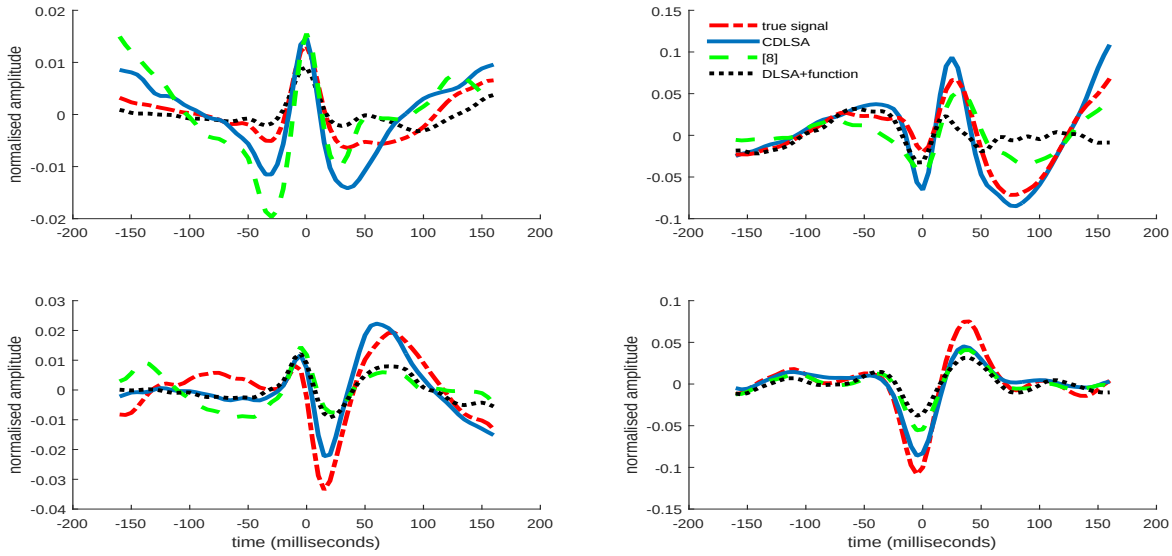

Fig. 3. Examples of single-channel reconstructed intracranial signals for four different subjects. Averaging the obtained time courses over trials we get waveforms for the true test signal, our method, the method in [8] and the DLSA+C method.

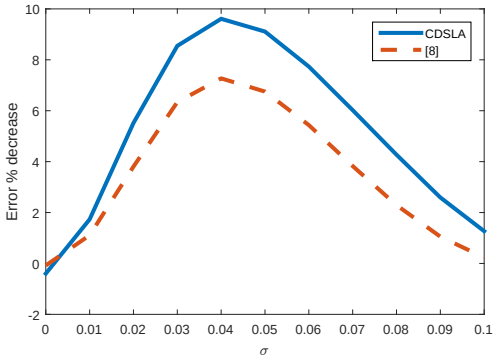

Fig. 2. Group average percentage decrease in error on the test set between our CDLSA method and the method in [8] with the DLSA+C method described in Section 2.3. The mapping $\mathbf{C}$ was assumed known and was fixed for both our and the DLSA+C methods.

Table 1. Reconstruction error in the test set between the estimated and true intracranial signal. We compare our CDLSA method, the DLSA+C method described in Section 2.3 and the one in [8].

Subject/Test Error	CDLSA	DLSA+C	[8]
1	0.073	0.076	0.073
2	0.187	0.191	0.188
3	0.180	0.185	0.182
4	0.116	0.121	0.117
5	0.113	0.116	0.111
6	0.068	0.070	0.070
7	0.109	0.114	0.106
8	0.095	0.095	0.102
9	0.067	0.077	0.069
10	0.091	0.092	0.090

4. CONCLUSIONS

We developed an algorithm that estimates a coupled dictionary and a mapping function $\mathbf{C}$ for a concurrent iEEG and sEEG epileptic dataset. The dictionary and mapping function were used by scalp-only EEG segments in order to estimate which atoms best describe and model those segments. Subsequently, we transformed the solution to the corresponding intracranial signal. The algorithm was able to better reconstruct the intracranial signal compared to that done by traditional dictionary learning.

For simulated data our method performed equally well as a similar method in [8] when the mapping $\mathbf{C}$ was learned from the data. On the other hand, it performed better when the mapping $\mathbf{C}$ was assumed known and fixed. The main reason for this is that since in [8] the dictionaries of both modalities can take any form they are more prone to overfitting the noise. A known $\mathbf{C}$ overcomes that issue. For real epileptic data, both coupled dictionary methods performed equally well however our method enables the estimation of $\mathbf{C}$. For future work we will develop physiologically plausible solutions $\mathbf{C}$ so that we obtain better solutions than the approximation of Eq. (12), develop methods that allow different $\mathbf{C}$ to be used for different trials, and learn a coupled wavelet based dictionary.

5. REFERENCES

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

[2] P. Frossard, "What is the right representation for my signal?," *IEEE Signal Processing Magazine*, , no. 3, pp. 27–38, 2011.

[3] D. E. Carlson, J. T. Vogelstein, W. Qisong, L. Wenzhao, Z. Mingyuan, C.R. Stoetzner, D. Kipke, D. Weber, D.B. Dunson, and L. Carin, "Multichannel electrophysiological spike sorting via joint dictionary learning and mixture modeling," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 1, pp. 41–54, Jan 2014.

[4] S. Shapoori, S. Sanei, and W. Wang, "A novel approach for detection of medial temporal discharges using blind source separation incorporating dictionary look up," in *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, April 2015, pp. 894–897.

[5] R. Mehrotra, D. Chu, S. Haider, and I. Kakadiaris, "It takes two to tango : Coupled Dictionary Learning for Cross Lingual Information Retrieval," *NIPS 2012*, pp. 1–5, 2012.

[6] C. Reale and R. Chellappa, "Coupled dictionaries from thermal to visible face recognition," *IEEE ICIP*, pp. 3–7, 2014.

[7] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2216–2223, 2012.

[8] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, pp. 2861–2873, 2010.

[9] Y. Li and X. Feng, "Coupled dictionary learning method for image decomposition," *Science China Information Sciences*, vol. 56, no. 3, pp. 1–10, 2013.

[10] D. Baby, T. Virtanen, T. Barker, and H. Van hamme, "Coupled dictionary training for exemplar-based speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 2883–2887.

[11] Y. Li and A. Ngom, "Sparse representation approaches for the classification of high-dimensional biological data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 2, pp. 447–456, 2013.

[12] D. Nayak, A. Valentín, G. Alarcón, Jorge J. García S., F. Brunnhuber, J. Juler, C. E Polkey, and C. D. Binnie, "Characteristics of scalp electrical fields associated with deep medial temporal epileptiform discharges.," *Clinical neurophysiology*, vol. 115, no. 6, pp. 1423–35, June 2004.

[13] H. G. Wieser, C. E. Elger, and S. R. Stodieck, "The foramen ovale electrode: a new recording method for the preoperative evaluation of patients suffering from mesio-basal temporal lobe epilepsy," *Electroencephalogr Clin Neurophysiol*, vol. 61, pp. 314–22, 1985.