A CLUSTERING APPROACH TO CONSTRUCT MULTI-SCALE OVERCOMPLETE DICTIONARIES FOR ECG MODELING

David Meltzer and David Luengo

Universidad Politécnica de Madrid, 28031 Madrid (Spain)

ABSTRACT

The electrocardiogram (ECG) is the main biomedical signal used to diagnose and monitor cardiac pathologies. A typical ECG is composed of quasi-periodic activations (the QRS complexes, and the P and T waves) and periods of inactivity, plus noise and interferences. The sparse nature of the ECG has lead to the development of many compressed sensing (CS) and sparsity-aware ECG signal processing algorithms. In order to attain a good performance, these methods require appropriate dictionaries, and several on-line dictionary construction approaches have been devised. However, all of them require a substantial computational cost and the derived dictionaries are composed of atoms which may not be representative of real-world signals. In this work, we describe an efficient method for off-line construction of an overcomplete and multi-scale dictionary using a clustering-based approach. The resulting dictionary, whose atoms are the most representative waveforms from the training set, is then used to obtain a sparse representation of the ECG signal. Simulations on real-world records from Physionet's PTB database show the good performance of the proposed approach.

Index Terms— ECG signal processing, sparse inference, off-line dictionary learning, hierarchical clustering, LASSO

1. INTRODUCTION

The electrocardiogram (ECG) is the main biomedical signal used to ascertain the cardiovascular health status of many patients, both in clinical and ambulatory settings [1]. A typical ECG is composed of quasi-periodic activations (the QRS complexes, and the P and T waves) and periods of inactivity (i.e., isoelectric intervals, like PQ or ST segments), plus noise and interferences (baseline wander, AC interference, electromyographic noise, motion artifacts, etc.) [1, 2].

The sparse nature of the ECG has lead to the development of many compressed sensing (CS) and sparsity-aware ECG signal processing algorithms. In order to attain a good performance, these methods require appropriate dictionaries composed of atoms that properly represent the significant waveforms observed in the observed signals. Many approaches have been devised for the sparse representation of single-channel and multi-channel ECGs using different types of simple analytical waveforms: Gaussians [3, 4, 5], generalized Gaussians and Gabor dictionaries [6], several families of wavelets (like the mexican hat or the coiflet4) [7, 8], etc. Unfortunately, although these approaches can attain good results, the obtained representations usually include many spurious activations that must be removed, e.g., by performing a post-processing stage [4, 8] or through the minimization of a complex non-convex cost function [9].

Realizing that a customized dictionary, built from real-world signals, would provide a better performance in terms of the achieved reconstruction error for a given level of sparsity, several *on-line* dictionary learning algorithms (e.g., see [10, 11]) have been developed. Some of these approaches have been applied in the context of ECGs: the K-SVD algorithm in [12], the shift-invariant K-SVD in [13], and the method of optimal directions in [14]. However, these methods have a high computational cost (due to their need to iterate between the dictionary learning and sparse approximation stages) and lead to dictionaries whose atoms may not be representative of real-world signals (thus reducing the interpretability of the sparse model, as well as the ability to easily locate the relevant waveforms). Alternatively, an *off-line* dictionary construction methodology (where a dictionary with real-world waveforms is initially built and then directly used for CS and sparse modeling without any further modification) has been proposed by Fira *et al.* [7, 15, 16]. However, in these approaches the atoms of the dictionary are either selected randomly from segments of the signal or taken directly from the first half of the ECG without any waveform selection.

In this work, we describe an efficient method to construct an overcomplete and multi-scale dictionary for sparse ECG representation using waveforms recorded from real-world patients. Unlike on-line dictionary learning methods, we first learn the dictionary off-line, and then we apply an efficient sparse inference algorithm (CoSa [17]) to model the signal using the learnt dictionary. As a result, our method is much more efficient from a computational point of view than other existing methods, thus becoming amenable to deal with long recordings from multiple patients. With respect to the approach of Fira *et al.*, our method selects

Thanks to Ministerio de Economía y Competitividad of Spain for funding through the MIMOD-PLC project (ref. TEC2015-64835-C3-3-R).

the optimal atoms to construct the dictionary, thus resulting in a much more compact solution. Numerical simulations demonstrate that the proposed approach is able to obtain a very sparse representation without missing any QRS complex or introducing spurious activations.

2. DICTIONARY CONFORMATION

The main goal of this work is to derive an efficient method to construct an overcomplete and multi-scale dictionary whose atoms are recorded QRS complexes extracted from actual patients.[1] In order to achieve this goal, a selection of the most representative QRS complexes available in the training set has to be performed in such a way that the selected waveforms represent disjoint subsets of meaningful and similar QRS complexes. In the following, we briefly describe the database used and the pre-processing steps performed, before detailing the off-line clustering-based approach followed to construct the dictionary.

2.1. Database

In order to construct the dictionary, we use the Physikalisch-Technische Bundesanstalt (PTB) database, compiled by the National Metrology Institute of Germany for research, algorithmic benchmarking and teaching purposes [18]. The ECGs were collected from healthy volunteers and patients with different heart diseases by Prof. Michael Oeff, at the Dep. of Cardiology of Univ. Clinic Benjamin Franklin in Berlin (Germany), and can be freely downloaded from Physionet [19].[2] The database contains 549 records from 290 subjects (aged 17 to 87 years) composed of 15 simultaneously measured signals: the 12 standard leads plus the 3 Frank lead ECGs [1, 2]. Each signal is digitized using a sampling frequency $f_s = 1000$ Hz with a 16 bit resolution. Out of the 268 subjects for which the clinical summary is available, we selected channel 10 (lead V4) of the first recording of the 52 healthy patients available in order to build the dictionary.

2.2. Pre-processing: Identification of QRS complexes

The pre-processing step consists in the extraction of all the QRS complexes from each ECG in the training set. Firstly, we apply a 4th order Butterworth bandpass filter with cut-off frequencies $f_{c1} = 1$ Hz and $f_{c2} = 40$ Hz to remove noise and interferences. Forward-backward filtering, with an appropriate choice of the initial state to remove transients, is used to avoid phase distortion. In a second stage, each ECG is processed for R-peak identification using the well-known Pan-Tompkins algorithm [20]. Then, each R-peak detected is provided to an algorithm that, by previously extracting a number of different fiducial points [21], determines the start (μ) and end (j-point) of the actual QRS complex by finding their location using the minimum radius of curvature.

The extracted μ to j-point QRS complexes are described by a similar, but variable, number of samples. In order to facilitate the comparison of the available QRS complexes using the distance metric described in Section 2.3, an equal number of samples is required for all of them. A common approach is to select each R-peak and take into consideration an equal number of samples on its left and right sides. This method can even eliminate the need for resampling, provided an equal sampling rate has been used for all the available signals. A drawback for this approach is precisely the simplifying (implicit) assumption that a QRS complex is centered on the R-peak or, in other words, that the Ventricular Activation Time represents half of the QRS duration. An alternative method that centers each QRS complex estimation in the R-peak but does not consider an equal number of original samples on its left and right sides has been described in [16]. It separately resamples the left and right R-peak sides to achieve an homogeneous number of samples. However, this method does not guarantee that the actual ratios of the different waveforms and QRS complex durations are maintained after the resamplings.

The approach followed in this work is based on resampling each QRS complex (from μ to j-point) by a variable factor L/M, which depends on its length, to achieve a constant final number of samples equal to the longest QRS complex identified [22, 23]. This approach guarantees that the aforementioned ratios are maintained after resampling, thus enabling the distance metric to quantify even the fact that waves can have different durations. Special care has been taken when resampling sequences to avoid edge effects produced by the antialiasing filter, given that the QRS complexes generally do not start and finish with zero values. A simple and effective method has been implemented for this purpose:

1. For each QRS complex, $x[n]$ for $n = 0, 1, \ldots, N-1$, two new sequences are constructed:

$$x_\ell[n] = x[n] - x[0], \qquad x_r[n] = x[n] - x[N-1].$$

2. The rational resampling factor for the QRS complex in order to achieve the desired length, $\frac{M}{L}$, is computed and the two sequences $x_\ell[n]$ and $x_r[n]$ are resampled to obtain the sequences $\tilde{x}_\ell[n]$ and $\tilde{x}_r[n]$.

3. Since $\tilde{x}_\ell[n]$ is not affected by the edge effect on its left-hand side, whereas $\tilde{x}_r[n]$ lacks any edge effect on its right-hand side, the resampled QRS complex is finally

$$\tilde{x}[n] = \begin{cases} \tilde{x}_\ell[n], & 0 \le n \le \left\lfloor \frac{N}{2}\frac{L}{M} \right\rfloor - 1; \\ \tilde{x}_r[n], & \left\lfloor \frac{N}{2}\frac{L}{M} \right\rfloor - 1 < n \le \left\lfloor N\frac{L}{M} \right\rfloor - 1. \end{cases}$$

[1] Let us remark that we focus here on the QRS complexes because they are the most relevant waveforms that can be found in the ECGs. However, the proposed approach can also be applied to construct dictionaries of typical P and T waveforms, as well as combined P-QRS-T dictionaries.

[2] PTB database: `https://www.physionet.org/physiobank/database/ptbdb/`

2.3. QRS complexes similarity metric

In order to perform the clustering of the available QRS complexes, a similarity metric must be selected first. Different quantifications of similarity measures for discrete-time sequences, like the inverse of the distance or the cross-correlation, can be applied. In this work, we have used the squared Euclidean distance (i.e., the energy of the difference among the two QRS complexes) to construct the clusters. Mathematically, given two discretized and equal length signals, $x[n]$ and $y[n]$ for $n = 0, 1, \ldots, N-1$, the distance expression is

$$d(x[n], y[n]) = \|\mathbf{x} - \mathbf{y}\|_2^2 = \sum_{n=1}^{N} (x[n] - y[n])^2.$$

In order to obtain a proper distance metric, the QRS complexes considered in this work have been previously energy normalized. For each discrete-time resampled QRS complex, $\tilde{x}[n]$, the energy normalized QRS complex $\bar{x}[n]$ is given by

$$\bar{x}[n] = \frac{\tilde{x}[n]}{\|\tilde{\mathbf{x}}\|_2} = \frac{\tilde{x}[n]}{\sqrt{\sum_{n=1}^{N} \tilde{x}[n]^2}}.$$

2.4. Grouping similar QRS complexes

Many different methods to group similar elements of a set into classes when a similarity metric is defined have been proposed in the literature. One of the main classes of methods is based on clustering algorithms. Among all the available clustering algorithms, we have decided to use *hierarchical agglomerative clustering* techniques, which require a lower computational effort than divisive techniques and thus result in more efficient methods. This group of techniques is based on the use of a proximity matrix that quantifies the similarity or dissimilarity of each pair of identified clusters in order to determine cluster aggregations in each iteration of the algorithm. A proximity matrix can be defined as a square $R \times R$ matrix, $\mathbf{D}$, whose (i, j)-th element contains the distance (dissimilarity) $d_{i,j} = d(C_i, C_j)$ among each pair of clusters C_i and C_j for $1 \leq i, j \leq R$:

$$D = \begin{pmatrix} 0 & d_{1,2} & \cdots & d_{1,R} \\ d_{2,1} & 0 & \cdots & d_{2,R} \\ \vdots & \vdots & \ddots & \vdots \\ d_{R,1} & d_{R,2} & \cdots & 0 \end{pmatrix}$$

Agglomerative clustering is performed using the iterative approach described in Algorithm 1.

A number of different agglomerative hierarchical clustering algorithms exist, depending on how the distance metric between clusters is defined. In this work, tests have been performed using single, complete, centroid and Ward's linkage methods. One important issue regarding classic hierarchical clustering methods is their lack of robustness to outliers. The performed experiments using different methods to calculate the distance among clusters show that outliers tend to be kept as singleton clusters for the single and centroid methods, while the rest of the mentioned methods do not show experimental issues related to outliers. However, a meaningful dictionary can still be constructed by discarding the singleton clusters, which can be easily identified. Figure 1 shows the graphs corresponding to the case where 2 clusters have been identified by means of hierarchical agglomerative clustering using the single linkage method. For each cluster, all of its elements (QRS complexes) are depicted, altogether with the representative element obtained as described in Section 2.5.

Start with R singleton clusters;
Calculate the proximity matrix for R clusters;
repeat
  Search for a pair of clusters C_i and C_j with the minimal distance $d_{i,j} = \min_{\substack{1 \leq m,l \leq R \\ m \neq l}} d(C_m, C_l)$;
  Join clusters C_i and C_j to form a new cluster;
until *the desired number of clusters remains*;

Algorithm 1: Hierarchical agglomerative clustering.

2.5. Selecting a representative QRS complex per cluster

Let us recall that the main goal of the proposed technique is obtaining a dictionary of QRS complexes directly extracted from recorded ECGs. Hence, once the desired number of clusters has been identified, a single QRS complex from each cluster has to be selected to conform the dictionary. As not all the methods for hierarchical clustering use the concept of cluster center, and even then the cluster's center does not necessarily correspond to an element of the cluster, the *medoid* has been chosen as the representative QRS complex of each cluster. If a cluster C includes P elements $(c_1, c_2, ..., c_P)$ and a distance metric $d(c_i, c_j)$ is defined, the medoid of the cluster is an element $c_{medoid} \in C$ such that

$$c_{medoid} = \underset{y \in \{c_1, c_2, \ldots, c_P\}}{\arg\min} \sum_{p=1}^{P} d(y, c_p)$$

Figure 1 shows in its upper graph the representative QRS complexes selected as medoids for each cluster. The middle and lower graphs show the superposed signals belonging to each cluster, altogether with their representative QRS complex (i.e., their medoids).

3. NUMERICAL RESULTS

In this section, we investigate the performance of the proposed clustering-based approach to construct the dictionary. In

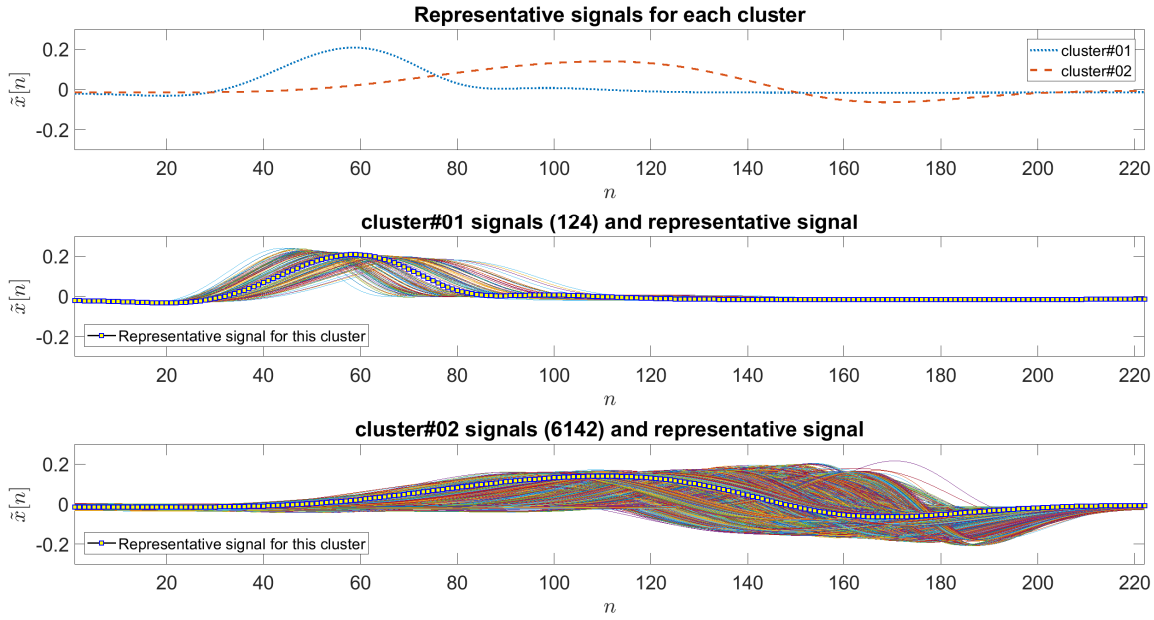

Fig. 1. Two clusters with their associated QRS complexes and their representative QRS complexes (i.e., the cluster's medoids).

Table 1. Coefficient sparsity, compression ratio (CR) and reconstruction SNR (R-SNR) vs. λ and cluster number (K).

	$K=1$			$K=2$			$K=3$			$K=4$		
λ	**Sparsity**	**CR**	**R-SNR**	**Sparsity**	**CR**	**R-SNR**	**Sparsity**	**CR**	**R-SNR**	**Sparsity**	**CR**	**R-SNR**
1	66.26%	0.36	18.2 dB	81.12%	0.33	22.2 dB	87.34%	0.32	22.3 dB	90.44%	0.34	21.8 dB
2	76.76%	0.59	15.7 dB	87.64%	0.57	18.5 dB	91.78%	0.56	18.6 dB	93.77%	0.60	17.3 dB
5	87.50%	1.46	11.3 dB	93.80%	1.74	12.6 dB	95.89%	1.69	12.6 dB	96.98%	1.91	11.1 dB
10	93.65%	3.09	7.7 dB	96.94%	4.82	8.3 dB	97.99%	4.32	8.4 dB	98.67%	4.48	7.4 dB
20	97.53%	9.13	4.5 dB	98.88%	9.25	4.9 dB	99.23%	8.10	4.9 dB	99.45%	7.93	4.5 dB

particular, our main goal is determining the quality of the sparse reconstruction when the dictionary is built from a set of patients and then used to model a different set of individuals (unlike most on-line approaches, which build a personalized dictionary for each signal). In order to achieve this goal, we apply the pre-processing described in Section 2.2 (bandpass filtering and QRS extraction) to the first register from all the healthy subjects (52) in the PTB database: patients 104, 105, 116, 117, 121, 122, 131, 150, 156, 166, 169, 170, 172–174, 180, 182, 184, 185, 198, 214, 229, 233–248, 251, 252, 255, 260, 263, 264, 266, 267, 276, 277, 279, and 284. Out of those 52 patients, the fiducial point extraction fails in three cases (patients 173, 245 and 284) and a total of 6266 QRS complexes ($\approx$ 127.9 QRS complexes per patient, with up to 194 QRS complexes for a single patient) are correctly extracted from the remaining 49 registers. This is the training set used to construct the dictionary as described in Algorithm 1 with a variable number of clusters from 2 to 4.

In order to test the proposed approach, we build a test set composed of all the available signals from those three challenging patients where we have been unable to properly locate the QRS complexes. Since two recordings are available for patient 245 and three recordings have been performed for patient 284, our test set contains 6 registers. A sparse approximation of channel 10 in these 6 recordings is performed using the LASSO. The recently proposed CoSa algorithm, using several values of the sparsity parameter λ (the higher the value of λ the sparser the solution), is used to obtain an efficient solution of the reconstruction problem [17]. Table 1 shows the average sparsity attained in the coefficients, the compression ratio (CR) and the reconstruction signal to noise ratio (R-SNR) as a function of the sparsity factor (λ) and number of clusters (K). These results show the feasibility of using a single dictionary in order to model multiple patients.

4. CONCLUSIONS

A novel, efficient, clustering-based off-line dictionary learning approach for ECG signals (QRS complexes) has been described in this paper. The performance of the proposed algorithm has been validated using signals from Physionet's PTB database, showing that effective sparse representations of multiple patients can be attained using a single dictionary. Future lines include developing mega-dictionaries with other waveforms (e.g., P and T waves), including patients with pathologies, and learning the optimal number of clusters.

5. REFERENCES

[1] L. Sörnmo and P. Laguna, *Bioelectrical signal processing in cardiac and neurological applications*, Academic Press, 2005.

[2] G.D. Clifford, F. Azuaje, and P.E. McSharry, Eds., *Advanced methods and tools for ECG data analysis*, Artech House, 2009.

[3] M.S. Billah, T.B. Mahmud, F.S. Snigdha, and M.A. Arafat, "A novel method to model ECG beats using Gaussian functions," in *Proc. 4th IEEE Int. Conf. on Biomedical Engineering and Informatics (BMEI)*, 2011, vol. 2, pp. 612–616.

[4] S. Monzón, T. Trigano, D. Luengo, and A. Artés-Rodríguez, "Sparse spectral analysis of atrial fibrillation electrograms," in *Proc. IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2012, pp. 1–6.

[5] T. Trigano, V. Kolesnikov, D. Luengo, and A. Artés-Rodríguez, "Grouped sparsity algorithm for multichannel intracardiac ECG synchronization," in *Proc. 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 1537–1541.

[6] O. Divorra-Escoda et al., "Ventricular and atrial activity estimation through sparse ECG signal decompositions," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2006, pp. 1060–1063.

[7] M. Fira, L. Goras, C. Barabasa, and N. Cleju, "On ECG compressed sensing using specific overcomplete dictionaries," *Advances in Electrical and Computer Engineering*, vol. 10, no. 4, pp. 23–28, 2010.

[8] D. Luengo, S. Monzón, T. Trigano, J. Vía, and A. Artés-Rodríguez, "Blind analysis of atrial fibrillation electrograms: a sparsity-aware formulation," *Integrated Computer-Aided Engineering*, vol. 22, no. 1, pp. 71–85, 2015.

[9] D. Luengo, J. Vía, S. Monzón, T. Trigano, and A. Artés-Rodríguez, "Cross-products LASSO," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6118–6122.

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

[11] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27–38, 2011.

[12] C. Wang, J. Liu, and J. Sun, "Compression algorithm for electrocardiograms based on sparse decomposition," *Frontiers of Electrical and Electronic Engineering in China*, vol. 4, no. 1, pp. 10–14, 2009.

[13] B. Mailhé et al., "Dictionary learning for the sparse modelling of atrial fibrillation in ECG signals," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 465–468.

[14] L.F. Polania and K.E. Barner, "Multi-scale dictionary learning for compressive sensing ECG," in *Proc. of the IEEE Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE)*, 2013, pp. 36–41.

[15] M. Fira, L. Goras, C. Barabasa, and N. Cleju, "ECG compressed sensing based on classification in compressed space and specified dictionaries," in *Proc. of the 19th European Signal Processing Conference (EUSIPCO)*, 2011, pp. 1573–1577.

[16] M. Fira and L. Goras, "On projection matrices and dictionaries in ECG compressive sensing — a comparative study," in *Proc. 12th Symp. on Neural Network Applications in Electrical Eng. (NEUREL)*, 2014, pp. 3–8.

[17] T. Trigano, Igor Shevtsov, and D. Luengo, "CoSA: An accelerated ISTA algorithm for dictionaries based on translated waveforms," *Signal Processing*, vol. 139, pp. 131–135, 2017.

[18] R. Bousseljot, D. Kreiseler, and A. Schnabel, "Nutzung der EKG-signaldatenbank CARDIODAT der PTB über das Internet," *Biomedizinische Technik/Biomedical Engineering*, vol. 40, no. s1, pp. 317–318, 1995.

[19] A.L. Goldberger et al., "Physiobank, physiotoolkit, and physionet," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[20] J. Pan and W.J. Tompkins, "A real-time QRS detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 32, no. 3, pp. 230–236, 1985.

[21] S.A. Israel et al., "ECG to identify individuals," *Pattern Recognition*, vol. 38, no. 1, pp. 133–142, 2005.

[22] D. Luengo, D. Meltzer, and T. Trigano, "Sparse ECG representation with a multi-scale dictionary derived from real-world signals," in *Proc. of the 41st Int. Conf. on Telecomm. and Signal Proc. (TSP)*, 2018, pp. 1–5.

[23] D. Luengo, D. Meltzer, and T. Trigano, "An efficient method to learn overcomplete multi-scale dictionaries of ECG signals," *Applied Sciences*, vol. 8, no. 12, pp. 2569–2586, 2018.