Long-Term ECG monitoring with zeroing Compressed Sensing approach

This is the author's version of the article that has been presented at IEEE NORCAS2015
The editorial version of the paper is available at http://dx.doi.org/10.1109/NORCHIP.2015.7364394

# Long-Term ECG Monitoring with Zeroing Compressed Sensing Approach

M. Mangia[‡], D. Bortolotti[†], A. Bartolini[†§], F. Pareschi[*], L. Benini[†§], R. Rovatti[†‡] and G. Setti[*‡]
[†]DEI, [‡]ARCES, University of Bologna, Italy - Email: {daniele.bortolotti, mauro.mangia2, riccardo.rovatti}@unibo.it
[§]Integrated Systems Laboratory, ETH Zurich, Switzerland - Email: {barandre, lbenini}@iis.ee.ethz.ch
[*]ENDIF, University of Ferrara, Italy - Email: {fabio.pareschi, gianluca.setti}@unife.it

***Abstract*—Novel low-voltage, low latency, non-volatile memory (NVM) technologies allow long-term wearable biomedical monitors to benefit from large storage capability, avoiding costly wireless transmissions and enabling, along with proper signal processing and architectural optimization, minimal energy operations and extended battery life. The recently proposed rakeness-based Compressed Sensing (RCS) offers high compression rate with an associated low computational power. This allows an energy trade-off between the compression stage and the storage stage. In this paper we introduce a novel approach, namely *zeroing* CS, which reduces RCS computational requirements to extremely low levels. The new energy trade-off is analyzed, considering a suitable multi-core DSP and different NVM technologies for local storage. According to our analysis, the proposing zeroing approach is up to 80% more efficient than a standard CS solution and 70% w.r.t. RCS when overall energy requirement is not dominated by storage.**

## I. Introduction

Human modern behavior-related diseases such as cardiovascular ones require continuous and long-term medical supervision, with increasing and unsustainable costs for the traditional healthcare system [1]. A scalable and cost-effective solution to this problem is offered by ultra-low power (ULP) personal monitoring systems (PMS), enabling ubiquitous and long-term monitoring policies for the future healthcare system.

However, the design of such devices is subject to two conflicting requirements: the power budget must be reduced for an extended battery life, while high computation capabilities are required to process and reduce the amount of data to be locally stored for further medical analysis. In this direction, the Compressed Sensing (CS) paradigm [6] has shown to be a very good candidate for lowering power requirements with respect to state-of-the-art compression algorithms [7], especially in electrocardiogram (ECG) signals compression [8], [9]. Recently, the standard CS (SCS) theory has been enhanced by the concept of *rakeness* [12], that exploits statistical features of the input signal (i.e., *localization*) to further increase the achievable compression rate at a given reconstruction quality level, and thus determining an additional reduction in terms of energy requirements in a CS-based system [13].

To find the right trade-off between PMS size, energy and storage, several aspects must be considered. Using novel non-volatile memory technologies, such as ReRAM [2], PCM [3], STT-MRAM [4] in the design of an ULP digital signal processor (DSP) for biomedical monitoring [5], may increase density level and lower access energy cost to a level that paves the way to new CS approaches that trade-off more carefully computational cost *vs* compression capabilities. In other words, when cost of storage is high and space is limited, then the most effective compression feasible for the DSP capabilities may be pursed. Complementary, if write access energy is small and memory capacity is large, then it may be convenient to reduce DSP activity (i.e., DSP energy consumption) at the cost of compression effectiveness.

The aim of this work is to address this opportunity by formalizing the zeroing CS (ZCS) technique that further reduces the energy requirement for processing. The basic idea is to randomly remove many mathematical operations characterizing the rakeness-based CS (RCS). This limits the compression capabilities with respect to RCS but at the same time outperforms SCS and RCS approaches in terms of memory footprint and energy requirements.

Motivated by the inherent parallel nature of medical-grade monitoring, a suitable multi-core DSP architecture is considered, as it proved its efficiency compared to single-core solutions [14]. Testing the proposed algorithmic solutions on such target architecture we show that: (*i*) RCS leads, with respect to SCS, to a $\approx 33\%$ improvement in terms of computational efficiency with an output data size reduction of $\approx 40\%$; (*ii*) the novel ZCS introduces several trade-offs in terms of input data compression, reconstruction quality, computational requirements and memory footprint. This approach proved its effectiveness for forth-coming NVM technolgies with respect to both SCS and RCS, with up to $\approx 80\%$ energy savings compared with SCS, and 70% w.r.t. RCS when overall energy in not dominated by storage.

The rest of the paper is organized as follows. In Section II concepts of CS, rakeness and zeroing are introduced. Section III presents the overall monitoring system architecture and the description of the experimental setup and the results in terms of energy efficiency and monitoring time considering different technological solutions for output storage. Conclusions are finally drawn.

## II. CS based on Rakeness and Zeroing

Let us refer to a time-limited discrete-time signal $x \in \mathbb{R}^N$, defined by the $N$ Nyquist-rate samples of the original analog signal $x(t)$. CS aims to overcome the limit imposed by the Nyquist-Shannon sampling theorem under the assumption that each signal instance $x$ has a *sparse* representation. This means

This is the author's version of the article that has been presented at IEEE NORCAS2015
The editorial version of the paper is available at http://dx.doi.org/10.1109/NORCHIP.2015.7364394

that there is an $N$-dimensional *sparsity basis* $\Psi \in \mathbb{R}^{N\times N}$, such that for any $x$ we have $x = \Psi\alpha$, where $\alpha$ has at most $K \ll N$ non-zero components.

Given a *sensing matrix* $\Phi \in \mathbb{R}^{M\times N}$, $M < N$, the CS theory ensures [6] that the information content of $x$ is preserved in a $M$-dimensional measurement vector $y \in \mathbb{R}^M$, obtained by projecting $x$ on the $M$ rows of $\Phi$, i.e.:

$$y = \Phi x + \nu = \Phi\Psi\alpha + \nu \tag{1}$$

where $\nu$ is an additive term used to model non-idealities such as the quantization error or the input noise.

Since $\Phi\Psi \in \mathbb{R}^{M\times N}$, with $M < N$, the inversion of (1) to obtain the reconstructed signal $\hat{x}$ is an ill-posed problem with an infinite number of solutions. The impasse is overcome by looking for the sparsest $\hat{x} = \Psi\hat{\alpha}$, i.e. by determining

$$\hat{\alpha} = \arg\min_{\alpha} \|\alpha\|_{l_1} \;\text{ s.t. }\; \|\Phi\Psi\alpha - y\|_{l_2} < \epsilon \tag{2}$$

where $\|\cdot\|_{l_1} = \sum|\cdot|$ and $\|\cdot\|_{l_2} = \sum\cdot^2$ are the standard $l_1$ and $l_2$ norms and $\epsilon$ bounds the effects of the noise term $\nu$.

The convergence of $\hat{x}$ to $x$ is guaranteed [6] when $M \geq O(K\log(N/K))$ and $\Phi$ is made by independent and identically distributed (i.i.d.) random variables [10],[11], including a sequence of random binary antipodal symbols with equal probability to be $-1$ or $+1$ [12].

The CS theory has recently been expanded with the introduction of the concept known as *rakeness* [12], that allows either to increase CS reconstruction quality or to reduce $M$ (i.e. increase compression) at a given performance level. The basic idea behind this approach is to exploit *localization* of signals, i.e. the assumption that the information is not equally distributed in the whole domain, but that some realizations of the input process have a higher probability with respect to all other ones [12]. The goal is to collect ("rake") the maximum amount of energy by a statistical matching between $x$ and the $j$-th row $\phi_j$ of $\Phi$, while at the same time preserving the randomness of $\Phi$. More formally, let us model $\phi_j$ and $x$ as realizations of two stochastic processes $\underline{\phi}$ and $\underline{x}$. We can so define *rakeness* as $\rho(\underline{\phi},\underline{x}) = \mathbf{E}_{\underline{\phi},\underline{x}}\left[|\langle\phi_j, x\rangle|^2\right]$[1]. Hence, referring without loss of generality to a random antipodal sensing matrix $\Phi \in \{+1,-1\}^{M\times N}$, the idea of increasing the collected energy, under the constraint that the $\phi_j$ are random enough to preserve the reconstruction ability, is translated into the following optimization problem:

$$\max_{\underline{\phi}} \rho(\underline{\phi},\underline{x}) \quad \text{s.t. } \langle\phi_j,\phi_j\rangle = N \;\text{ and }\; \rho(\underline{\phi},\underline{\phi}) \leq \tau N^2 \tag{3}$$

where $\tau N^2$ is an upper bound of the randomness[2] of the $\underline{\phi}$. The output of the optimization problem (3), analytically solved in [12], is the second-order statistical characterization of $\underline{\phi}$. The problem of generating random sequences with a prescribed second-order statistic is far from being trivial, since standard techniques (see f.i. [15]) are insufficient and linear probability feedback processes based architectures [16] need to be exploited.

[1] $\mathbf{E}_{\underline{\phi},\underline{x}}$ stands for the expected value with respect to both $\underline{\phi}$ and $\underline{x}$.

[2] The tuning of $\tau$ is not critical, as shown in [17].

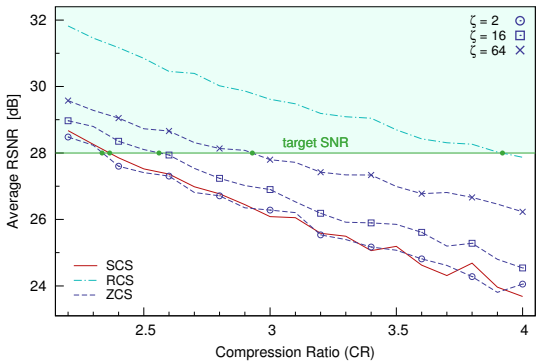


Fig. 1. Average RSNR as a function of the compression ratio (CR) for Standard CS, Rakeness-based CS and Zeroing CS with different $\zeta$ values.

When $\Phi$ is either an SCS or an RCS antipodal sensing matrix, the computational complexity of (1) is already low, since only $M \cdot N$ sums or subtractions are required. The basic idea underlying this paper is to further reduce this complexity with the ZCS approach, i.e. by *zeroing* some entry of $\Phi$ and allowing $\Phi \in \{+1, 0, -1\}^{M\times N}$. Starting from an antipodal $\Phi$, obtained with the RCS approach, we set to zero all $\Phi$ elements except $\zeta$ entries in each column, with $0 < \zeta \leq M$. The effect is a perturbation of the statistical characterization imposed by (3), with an expected reduction in performance, but also a reduction of the computational complexity of (1) to $\zeta N$ sums. This trade-off presents many advantages as shown in the following.

The SCS, RCS and ZCS approaches have been tested on real ECG signals from the MIT-BIH arrhythmia database [18]. For the sake of illustration, here we present results from 71.1 s of the the record 101. This signal is sampled at 360 Hz, and its signal-to-quantization noise ratio (SQNR) is estimated as 38.5 dB. The signal has been partitioned in 50 non-overlapping time windows with $N = 512$ samples each, and different values of $M$ are used. The considered sparsity matrix $\Psi$ is the Symmlet orthonormal basis. For each approach, a unique sensing matrix $\Phi$ has been selected for each value of $M$ by means of preliminary tests on synthetic ECGs [19]. More specifically, we generated 100 synthetic ECGs and: (*i*) for the SCS case, we chose $\Phi$ as the random antipodal matrix ensuring the best average reconstruction performance; (*ii*) for the RCS case, we used the same synthetic ECG generators to estimate the correlation profile required by (3) and we chose $\Phi$ again as the matrix guaranteeing best performance; (*iii*) for the ZCS case, we take the optimal matrix $\Phi$ in (*ii*) and randomly set to zero $M - \zeta$ entries in each of its column. This ensures the maximum fairness, since the $\Phi$ is not biased on any particular real signal.

The considered figure of merit is the reconstruction signal-to-noise ratio, defined as $\text{RSNR} = (\|x\|_{l_2}/\|x - \hat{x}\|_{l_2})_{\text{dB}}$, and averaged over the 50 considered time windows. The $\hat{x}$ is reconstructed by solving (2) (see [20] for more details). Results for different compression ratios $\text{CR} = N/M$, are shown in

This is the author's version of the article that has been presented at IEEE NORCAS2015
The editorial version of the paper is available at http://dx.doi.org/10.1109/NORCHIP.2015.7364394

TABLE I
CR, $M$ AND # OF SUMS NEEDED SNR=28 DB FOR VARIOUS CS CASES.

| CS | | CR | $M$ | # sums |
|---|---|---|---|---|
| SCS | | 2.36 | 217 | $1.1 \times 10^5$ |
| RCS | | **3.90** | **131** | $6.7 \times 10^4$ |
| ZCS | $\zeta = 64$ | 2.93 | 175 | $3.2 \times 10^4$ |
| | $\zeta = 16$ | 2.56 | 200 | 8192 |
| | $\zeta = 2$ | 2.34 | 219 | **1024** |

Fig. 1, where the SCS, the RCS and the ZCS approach with different values of $\zeta$ are considered. As expected, the RCS outperforms all other approaches, while performance for the ZCS is increasing with the $\zeta$ value, and drops to that of the SCS only for very low $\zeta$.

As an additional figure of merit, we consider the maximum CR for which system performance is assessed to a minimum target RSNR. In the example provided here, we set this threshold at 28 dB that represents in our experiments a good trade-off between the SQNR and a good visual representation of the signal. This is confirmed by Fig. 2 that shows short chunks of reconstructed ECG signals at the target RSNR. The maximum CR (and corresponding $M$) for all considered CS approaches, along with the computational complexity in terms of number of sums required to compute $y$, is shown in Table I.

With respect to the RCS, the zeroing technique can greatly reduce the computational cost required to compress a signal at the expense of a lower CR, i.e., of increasing $M$. As it will be shown in the following, in several cases this leads to a reduction in the overall energy requirement of the system. The optimal choice of $\zeta$ is therefore a trade-off depending on the power consumption of the DSP computing the CS encoding and on the energy required for storing the $M$ measurements.

## III. BIOMEDICAL MONITOR AND EVALUATION

The biomedical system we are considering, with its phases and architectural components, is shown in Fig. 3. It is characterized by a multi-core DSP architecture mimicking the solutions in [14], [13] which are particularly capable of targeting the digital biosignal processing domain. The architectural template features $P$ processing elements (PEs), which do not have instruction nor data caches, therefore avoiding refill costs and coherency protocol overheads. Each PE has a private

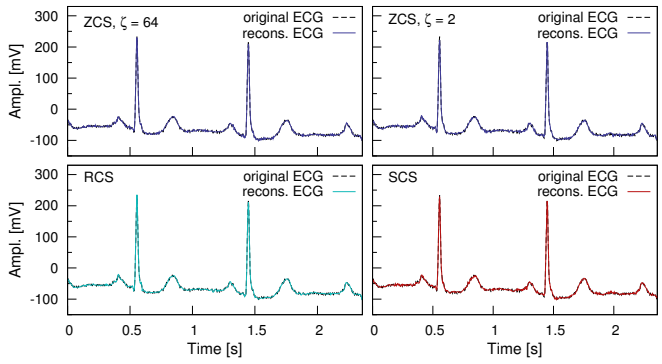


Fig. 2. Chunks of reconst. signal at the target RSNR for all CS algorithms.

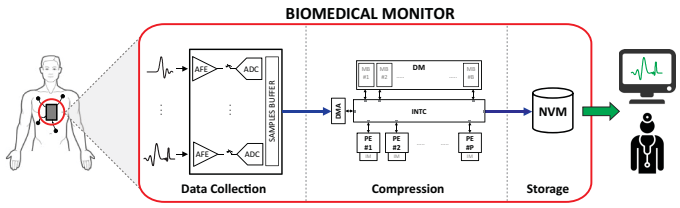


Fig. 3. Block scheme of the considered PMS and its architectural elements.

single-cycle instruction memory (IM) where the CS code is stored, while they all share a $B$-banks scratchpad data memory (DM) trough a suitable mesh-of-trees interconnection network (INTC), supporting single-cycle communication between PEs and memory banks [13]. During the *data collection* phase, the input $l$-channel biosignal is sampled by the Analog Front End (AFE) with a fixed sampling frequency ($f_s$), and the multi-core DSP waits for the samples required to perform compression. To reduce the AFE buffer size and avoid double buffering overhead, the DMA is triggered every $1/f_s$ seconds. Whenever a new set of samples is available, the DMA empties the AFE buffer and moves $l$ new samples into the DM, then once all the $N$ samples are copied the *compression* phase starts with the DSP performing the CS algorithm as a burst of computation on the available data[3]. The DSP is operating in a SIMD fashion, where each core is processing, in parallel with the others, the data relative to one of the $l$ input channels. After input data compression, the compressed output data are moved in a non-volatile memory (NVM) for future off-line medical analysis. During all the time where the DSP is idle, we assume a deep low-power state.

The considered multi-core DSP architectural template has been modeled and integrated in a SystemC-based cycle-accurate virtual platform [21], with back-annotated power numbers extracted from a RTL-equivalent architecture [22]. SystemC vs. RTL execution cycles misalignment is below 7%.

The 8-cores DSP architecture has been configured with a 16-banks DM, private IM of size 1 KB per core and a stack portion in DM of 512 B for each core (sufficient for the CS execution). Static data allocation is performed by means of cross-compiler attributes and linker script sections. From an algorithmic point of view, in both SCS and RCS the sensing matrix $\Phi$ is full and suitable for a normal vector-matrix projection. Instead if we consider the ZCS case, $\Phi$ is sparse and therefore suitable for a more efficient LUT-based algorithmic implementation [13]. Due to the varying amount of measurements required to achieve the target RSNR (see Table I) and to the different algorithmic implementations, the footprint in the DSP data memory varies. For all cases the DM memory has to allocate the input samples: considering $f_s = 360$ Hz, $N = 512$, $l = 8$ and 12-bit ADC resolution it leads to 6 KB. For the measurement vectors and the sensing matrix $\Phi$, in full or sparse form, the DM requirements are different: SCS = 111.9 KB, RCS = 67.5 KB and for the

[3]Note that in our architecture CS is used to compress signal in the digital domain (similar to what proposed in [7]) and not in the analog domain as originally considered in [6].

This is the author's version of the article that has been presented at IEEE NORCAS2015
The editorial version of the paper is available at http://dx.doi.org/10.1109/NORCHIP.2015.7364394

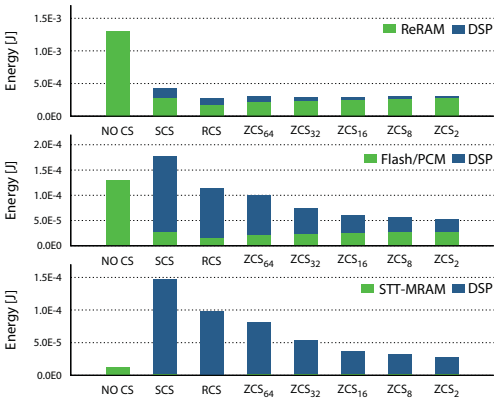


Fig. 4. Stacked bars of DSP compression energy and NVM storage energy for different NVM technologies, considering all CS approaches.

zeroing cases, $\text{ZCS}_{64} = 66.7\,\text{KB}$, $\text{ZCS}_{16} = 19.1\,\text{KB}$ and $\text{ZCS}_2 = 5.4\,\text{KB}$.

To evaluate the energy consumption of the compression stage we are considering a $28\,\text{nm}$ FD-SOI [23] technological library, a key technology to achieve ultra-low power operation for the CS workload requirements. More specifically, we consider the design corner (RVT,25 C,0.6 V) @ 10 MHz. To reduce power consumption, for the dynamic power the logic is clock-gated during the idle phases, while for leakage contribution we consider a reverse body bias voltage $\text{V}_{\text{RBB}} = 1\,\text{V}$, leading to $7\times$ leakage power reduction [23]. For what concerns the storage system we consider different NVM technologies, namely ReRAM, Flash/PCM and STT-MRAM. The energy required to store the data was assumed to be as follows: $\text{E}_{\text{ReRAM}} = 10\,\text{nJ/bit}$ [2], $\text{E}_{\text{Flash/PCM}} = 1\,\text{nJ/bit}$ [24], [3], $\text{E}_{\text{STT-MRAM}} = 0.1\,\text{nJ/bit}$ [4] and used as a parameter for the power model in [13]. As a comparison, during the compression phase the DSP in average consumes $\approx 3\,\text{nJ/bit}$.

The results are presented in Fig. 4. The plotted bars show the energy requirements within a time window for the DSP (blue), stacked with the energy required for storage in the NVM (green). For a fair comparison, we include the case of no compression, i.e. input data stored after sampling (NO CS).

For the DSP component, we can observe that RCS and ZCS always outperform SCS, where the RCS saves $\approx 33\%$ and for the zeroing approach the compression energy decreases, as expected, with the number of non-zeros, with the highest savings for $\text{ZCS}_2$ ($\approx 83\%$). Conversely, as the energy/bit decreases, ZCS gives greater benefits. Indeed for Flash/PCM, $\text{ZCS}_2$ achieves $\approx 70\%$ of energy savings w.r.t. standard CS (Fig. 4), while providing the same monitoring time (Table II). When the storage energy gets negligible, STT-MRAM in Fig. 4, two different trends can be observed: $\text{ZCS}_2$ achieves $81\%$ of energy savings w.r.t. standard CS, but loses $\approx 10\%$ w.r.t. no compression. Compared to no compression, $\text{ZCS}_2$ enables a 5x longer monitoring time. One can therefore conclude that RCS and ZCS allow to find the optimal design trade-off in between energy consumption and monitoring time.

TABLE II
MONITORING TIME AS A FUNCTION OF NVM CAPACITY.

| CS | 512 KB | 32 MB | 1 GB |
|---|---|---|---|
| | time ($m$) | time ($h$) | time ($d$) |
| NO CS | 0.76 | 0.81 | 1.08 |
| SCS | 3.58 | 3.82 | 5.09 |
| RCS | **5.93** | **6.32** | **8.43** |
| ZCS, $\zeta = 64$ | 4.44 | 4.73 | 6.31 |
| ZCS, $\zeta = 2$ | 3.55 | 3.78 | 5.04 |

## IV. CONCLUSION

RCS and its zeroing version enable trading off the computation requirements with the amount of data for later storage. In this paper we evaluated such trade-offs considering a multi-core DSP and different NVM technologies for storage. Experimental results showed that ZCS proves to be more energy efficient than the SCS approach when energy requirement for storage is significant. Moreover, when compression energy and storage energy are comparable, such approaches allow the flexibility of several design choices for what concerns energy consumption and monitoring time.

## REFERENCES

[1] World Health Organization [Online] http://www.who.int/mediacentre/factsheets/fs317
[2] M. Chang et al., *"A 0.5V 4Mb logic-process compatible embedded resistive RAM (ReRAM) in 65nm CMOS using low-voltage current-mode sensing scheme with 45ns random read time"*, ISSCC, 2012
[3] G. De Sandre et al., *"A 90nm 4Mb embedded PCM with 1.2V 12ns read access time and 1MB/s write throughput"*, ISSCC, 2010
[4] D. Halupka et al., *"Negative-resistance read and write schemes for STT-MRAM in 0.13m CMOS"*, ISSCC, 2010
[5] D. Son et al.,*"Multifunctional wearable devices for diagnosis and therapy of movement disorders"*, Nat. nanotech., v.9, pp.397–404, 2014
[6] E.J. Candes et al., *"An Introduction To Compressive Sampling,"* IEEE Signal Processing Mag., v.25, pp.21–30, 2008.
[7] H. Mamaghanian et al., *"Compressed Sensing for Real-Time Energy-Efficient ECG Compression on Wireless Body Sensor Nodes"*, IEEE T. on Biomedical Eng., v.58, pp.2456–2466, 2011
[8] Z. Zhilin et al., *"Compressed Sensing for Energy-Efficient Wireless Telemonitoring of Noninvasive Fetal ECG Via Block Sparse Bayesian Learning"*, IEEE T. on Biomedical Eng., v.60, pp.300–309, 2013
[9] A.S. Alvarado et al., *"Time-Based Compression and Classification of Heartbeats"*, IEEE T. on Biomedical Eng., v.59, pp. 1641–1648, 2012.
[10] S. Callegari et al., *"First direct implementation of a true random source on programmable hardware,"* Int. J. of Circ. Th. and App., v.33, pp. 1–16, 2005.
[11] F. Pareschi et al., *"Implementation and Testing of High-Speed CMOS True Random Number Generators Based on Chaotic Systems"*, IEEE Trans. Circuits Syst. I, vol. 57, no. 12, pp. 3124–3137, Dec. 2010.
[12] M. Mangia et al., *Rakeness in the design of analog-to-information conversion of sparse and localized signals*, IEEE T. on Circuits and Systems I, v.59, pp.1001–1014, 2012
[13] Bortolotti, D. et al., *"Rakeness-based Compressed Sensing on Ultra-Low Power Multi-Core Biomedical Processors"*, DASIP, 2014.
[14] A. Y. Dogan et al., *"Low-power processor architecture exploration for online biomedical signal analysis"*, IET Circuits, Devices & Systems, v.6, pp.279–286, 2012.
[15] M. Hasler et al., *"Special issue on applications of nonlinear dynamics to electronic and information engineering,"* Proceedings of the IEEE, v.90, pp.637–638, 2002
[16] R. Rovatti et al., *"Memory-antipodal processes: Spectral analysis and synthesis"*, IEEE T. on Circuits and Systems I, v.56, pp.156–167, 2009
[17] N. Bertoni, et al., *"Correlation tuning in compressive sensing based on rakeness: A case study"*, ICECS, 2013
[18] A.L. Goldberger et al., *"PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals"*, Circulation, v.101, pp. e215e220, 2000
[19] P. E. McSharry et al., *"A dynamical model for generating synthetic electrocardiogram signals"*, IEEE T. on Biomedical Eng., v.50, pp.289–294, 2003
[20] E. J. Candes, et al., *"Compressed sensing with coherent and redundant dictionaries"*, Applied and Comp. Harm. Anal., v.31, pp. 59–73, 2011
[21] D. Bortolotti et al., *"VirtualSoC: a Full-System Simulation Environment for Massively Parallel Heterogeneous System-on-Chip"*, IPDPWS, 2013
[22] M. Gautschi et al., *"Customizing an Open Source Processor to Fit in an Ultra-Low Power Cluster with a Shared L1 Memory"*, GLSVLSI 2014
[23] D. Jacquet et al., *"A 3 GHz Dual Core Processor ARM Cortex TM -A9 in 28 nm UTBB FD-SOI CMOS With Ultra-Wide Voltage Range and Energy Efficiency Optimization"*, IEEE J. of Solid-State Circuits, v.49, pp.812–826, 2014
[24] D. Shum et al., *"Highly Reliable Flash Memory with Self-Aligned Split-Gate Cell Embedded into High Performance 65nm CMOS for Automotive & Smartcard Applications"*, IMW, 2012.