Sequential Power per Area Optimization of Multichannel Neural Recording Interface Based on Dual Quadratic Programming

Amir Zjajo, Carlo Galuzzi, Rene van Leuken

Abstract—In this paper, we propose a novel method for power per area optimization under yield constrains in multichannel neural recording interface. Using a sequence of minimizations with iteratively-generated low-dimensional subspaces, our approach renders consistently improved power per area ratio and imposes no restrictions on the distribution of process parameters or how the data enters the constraints. The experimental results, obtained on neural recording interface circuits in CMOS 90nm technology, demonstrate power savings of up to 26% and area of up to 22% without yield penalty.

I. INTRODUCTION

The high density of neurons in neurobiological tissue requires a large number of recording electrodes in brain machine interface (BMI) to obtain accurate representation of the neural activity (e.g., for spatially broad analysis of neuronal synchronization), and to allow the location controllability of the recording sites [1]. Monitoring the activity of large number of neurons is a prerequisite for understanding the cortical structures and can lead to a better comprehension of severe brain disorders, such as Alzheimer's and Parkinson's diseases, epilepsy, autism and psychiatric disorders [2] or to reestablish sensory (e.g. vision, hearing) or motor (e.g. movement, speech) functions [3]. Multichannel electrode arrays in neuroprosthetic devices are combined with CMOS electronics for long-term, reliable, and stable recording of neural signals [4], on-chip processing of the recorded neural data [5], and stimulating the nervous system [6]. This migration, to allow proximity between electrodes and circuitry, and the increasing density in multichannel electrode arrays, are creating significant circuit design challenges with regard to miniaturization and power dissipation reduction. When integrating a large number of recording and stimulation channels on a single chip, low power dissipation becomes a major constraint even when they operate on a reliable power source. Power density is limited to 0.8 mW/mm^2 [7] to prevent possible heat damage to the tissue surrounding the device (and additionally to provide a longer battery life for implantable neuroprosthetic devices). Furthermore, the space to host the system is restricted to ensure minimal tissue damage and tissue displacement during implantation. As a consequence, intrinsic circuit noise is often traded for low power and high density of integration. Technology scaling, circuit topologies, architecture trends and post-silicon tuning

approaches specifically target power-performance trade-off. Circuit techniques such as current reuse [8], time multiplexing [9] and adaptive duty-cycling of the entire analog front end [10] can be used to improve power efficiency by exploiting the fact that neurons spikes are irregular and low frequency. Analytical optimization based on sensitivities [11] and physical [12] parameters offer guidelines for optimum power operation. The choice of the nonlinear optimization techniques including system-level hierarchical optimization [13], building-block-level optimization [14]-[15], and geometric programing [16] is based on the nonlinear relationships that exist between device lengths and widths and their associated performance due to strong short-channel effects in the nanometer region.

In this paper, we develop a yield constrained sequential power minimization framework based on dual quadratic program that is applied to multivariable optimization in neural interface design under bounded process variation influences. In the proposed algorithm, we create a sequence of minimizations of the feasible power per area ratio region with iteratively-generated low-dimensional subspaces, while accounting for the impact of area scaling. The proposed method can be used with any variability model, and is not restricted to any particular performance constraint. The yield constraint becomes active as the optimization concludes, eliminating the problem of overdesign in the worst case approach.

II. POWER PER AREA OPTIMIZATION OF MULTICHANNEL NEURAL RECORDING INTERFACE

A. Architectural Overview of a Multichannel Neural Recording Interface

With an increase in the range of applications and their functionalities, neuroprosthetic devices are evolving to a closed-loop control system [17] composed of a front-end neural recording interface and a back-end neural-signal processing, with features such as spike detection circuits [18] or LFP measurement circuits [19] for data reduction. The general BMI architecture includes, additionally, a micro-stimulation module to apply stimulation signals to the brain neural tissues. The block diagram of a *M*-channel neural recording system is illustrated in Figure 1. The data acquired by the recording electrodes is conditioned using analog circuits. As a result of the small amplitude of neural signals and the high impedance of the electrode tissue interface, low-noise amplification (LNA) and band-pass filtering of the neural signals is performed before the signals can be digitized by a successive approximation register (SAR)-based analog to digital converter (A/D converter).
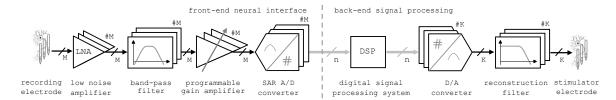
Figure 1: Block diagram of a brain machine interface with *M*-channel front-end neural recording interface and back-end signal processing.

To lower demands on driving capabilities of the amplifier and relax noise and cross-talk requirements, programmable gain amplifier (PGA) and SAR A/D converter are embedded in every recording channel. A low-power monolithic digital signal processing (DSP) unit provides additional filtering and executes a spike discrimination and sorting algorithms (to obtain data reduction and distinguish different neuronal sources). The relevant information is then transmitted to an outside receiver through the transmitter or used for *K*-channel stimulation in a closed-loop framework.

B. Circuit Parameters Formulation

The deterministic designable parameters d_r, $r = 1, ..., v_{ds}$ are denoted by the vector $d \in \mathcal{D}$, where $\mathcal{D}$ is the designable parameter space. The process parameters are treated as correlated random variables, whose means, standard deviations and correlation coefficients are obtained from process measurements as in [20]. We define yield as the percentage of manufactured circuits that meets all the specifications, considering process and environmental variations

$$y(d) = E\{ y(d, p_z) \mid pdf(p_z)\} \tag{1}$$

where $E\{.\}$ is the expected value and each vector d has an upper and lower bound determined by the technological process variation p_z with probability density function $pdf(p_z)$. Let the total area of the circuit be $A_{total} = \Sigma_k(x_k A_k)$, where A is the area of a transistor or a discrete component (resistor or capacitor), k is an index that runs over all transistors or a discrete components in the circuit and x is the sizing factor ($x \geq 1$). The optimization problem is then formulated as to find a design point that minimizes total power P_{total} over the deterministic designable parameters d with lower bounds a_j, and upper bounds b_j, for $1 \leq j \leq m$ in the design space $\mathcal{D}$, subject to a minimum yield requirement y with bound ξ

$$\min_{d \in \mathcal{D}(P_{total})} P_{total}(d)$$

subject to

$$a_j \leq d \leq b_j \quad 1 \leq j \leq m \tag{2}$$

$$y(d, p_z) \geq 1 - \xi \quad \forall d \in \mathcal{D}(P_{total})$$

$$x_k = 1 \quad \forall k \in \{1, 2, \ldots, q\}$$

Let $\mathcal{D}(P_{total})$ be the compact set of all valid design variable vectors d, such that $P_{total}(d) = P_{total}$. The designable parameter space $\mathcal{D}$ is assumed to be compact, which for all practical purposes is no real restriction when the problem has a finite minimum. The main advantage of this approach is its generality: it imposes no restrictions on the distribution of p and on how the data enters the constraints. If, as an approximation, we restrict $\mathcal{D}(P_{total})$ to just the one-best derivation of P_{total}, then we obtain the structured perceptron algorithm [21]. As a consequence, given active constraints, including optimum power budget and minimum frequency of operation, (2) can be effectively solved by a sequence of minimizations of the feasible region with iteratively-generated low-dimensional subspaces using a cutting plane method [22].

C. Power per Area Optimization

The power optimization problem involves varying the design point to optimize multiple performance objectives subject to constraints of other, secondary performance measures and designable parameter boundaries. With a metric power per area (PPA), we quantify the minimum power design that meets a targeted performance, while accounting for the impact of area scaling. The PPA metric depends on the technology node, process and operating conditions, circuit specification and the technology's V_T option. The PPA multi-criteria optimization problem is firstly translated into a min-max problem [23]. At any design point, the PPA value is converted into a performance score s. The individual performance scores s at a design point are used to compute an overall index of circuit quality, denoted by PPA $(d; s)$, which is the objective function for the design optimization. Thus, the constrained multi-criteria optimization problem is converted into an optimization problem with a single objective function. As a result, the general form of optimization problem becomes

$$\min_{d \in \mathcal{D}(P_{total})} PPA(d; s)$$

subject to

$$a_j \leq d \leq b_j \quad 1 \leq j \leq m \tag{3}$$

$$y(d, p_z) \geq 1 - \xi \quad \forall d \in \mathcal{D}(P_{total})$$

$$1 \leq x_k \quad \forall k \in \{1, 2, \ldots, q\}$$

To start the optimization problem, a design metric for global solution is initially selected, based on the priority given to the power budget as opposed to the performance function in a given application. If we assume that $\Delta(P_{total}, P_{total,i}) > 0$ for $i \in \{1, \ldots, N\}$, then the score s can be compactly written as a set of non-linear constraints

$$\forall i : \min_{d \in \mathcal{D}(P_{total})} \{PPA(d, \Psi(y_i, P_{total}))\} < PPA(d, \Psi(y_i, P_{total,i})) \tag{4}$$

where Ψ is a combined feature representation of a performance function in a given application. We replace each nonlinear inequality in (4) by $|\mathcal{D}|$-1 linear inequalities

$$\forall i, \forall P_{total} \in \mathcal{D} : PPA(d, \delta \Psi_i(P_{total})) > 0 \tag{5}$$

If the set of inequalities in (5) is feasible, typically there will be more than one solution d. For a unique solution, we select d with $\|d\| \leq 1$ for which s is uniformly different from the next closest score update. The score update is than expressed as dual quadratic program (QP)

$$
\max \quad -\frac{\eta}{2} \left\| \sum_{d' \in \mathcal{D}(P_{total})} \alpha_{d'} (h(d'_i) - h(d')) \right\|^2
$$
$$
+ \eta \sum_{d' \in \mathcal{D}(P_{total})} \alpha_{d'} PPA_i(d', d'_i; \delta \Psi_i(P_{total})) \quad (6)
$$
$$
\text{subject to} \quad \sum_{d' \in \mathcal{D}(P_{total})} \alpha_{d'} = 1 \qquad \alpha_{d'} \geq 0 \qquad \forall d' \in \mathcal{D}(P_{total})
$$

where η is the step size, $\alpha_{d'}$ the Lagrange multiplier enforcing the constraint for label $d' \neq d_i$. and $h(d)$ are the feature vectors of a design variable vector d. To find the local maxima and minima, we repeatedly select a pair of derivatives of d and optimize their dual (Lagrange) variables $\alpha_{d'}$. The dual program formulation has two important advantages over the primal QP; as dual program only depends on inner products defined by Ψ, it allows the usage of kernel functions and additionally, the constraint matrix of the dual program supports problem decomposition. At the end of sequence, we average all the score vectors s obtained at each iteration, similar to structured perceptron algorithm [21].

III. EXPERIMENTAL RESULTS

All the experimental results are carried out on a single processor Ubuntu Linux 9.10 system with Intel Core 2 Duo CPUs 2.66 GHz processor and 6 GB of memory. The circuit netlist is simulated in Cadence Spectre using 90nm CMOS model files. The simulation date points are processed with a PERL script and fed back into the MatLab code. The PPA ratio differs for each design depending on circuit characteristics, such as power consumption, bandwidth, gain, linearity, etc. Closed-form symbolic expressions of the constraints and the objective are passed on to the optimization algorithm. Design heuristics are used to provide a good initial starting point. The total run-time of the optimization method is only dozens of seconds, and the number of iterations required to reach the stopping criterion never exceeds 6 throughout the entire simulated β range (from 10^{-3} to 10^{-1}).

The design trade-off exploration space for circuit area, sample frequency and PPA is illustrated in Figure 2. The area and sample frequency curves are plotted for the worst-case design (WCD), and the proposed quadratic program optimized approach (QPO). The iso-PPA curves are plotted as overlay; the intersection with the area-sample frequency curves represents the normalized PPA ratio of the design. For a given circuit area, the optimized design obtains higher performance than the corresponding WCD. Alternatively, the optimized design assists in lower area designs for a given sample frequency. The points lying on the lowest intersections are most power efficient for the given input and output constraints and represent the PPA curve of interest. Power per area optimization for a fixed input size and output load constraint is the most common design scenario. The plot

in Figure 3 illustrates the position of the optimal power per area under maximum yield reference design point. In Table 1, the worst-case design (WCD) is compared across the neural interface circuits with the optimization approach. The QP optimized circuits allow large area reduction when designed for maximum WCD frequency ranging from 9% to 19%, with 16% on average. When operating at the same frequency, optimized total power is reduced up to 21%. In symmetrical circuit structures, the optimization space is restricted and, therefore, the additional power saving contributed by an optimization is much smaller, especially with the higher yield. For decreased yield, 95% instead of 99%, higher power saving of up to 32% on average can be achieved as a consequence of a larger optimization space (not shown in Table I). Note that over-dimensioning in a case of higher yield, leads to a larger area and higher power consumption.

The reduction of area for analog designs usually implies a trade-off, of which the most common is an increase in noise. Fortunately, the interface's input equivalent noise voltage decreases as the gain across the amplifying stages increases. The only noise sources in the LNA and g_m-C filter are the channel thermal noise of the transistors that make up the transconductor and the thermal noise of any degeneration resistor that are used for linearization of the transconductor. The observed circuit's power consumption scales with its bandwidth and signal-to-noise ratio (SNR). This lower bound on the speed is primarily a function of the technology's gate delay and kT/C noise multiplied by the number of SAR cycles necessary for one conversion.
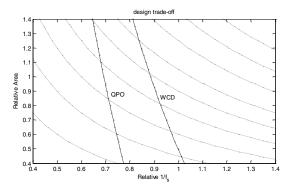

Figure 2: Area, sampling frequency and PPA trade-off for neural recording channel optimized with quadratic programming (QPO) and worst-case design (WCD). PPA is shown as an overlay.
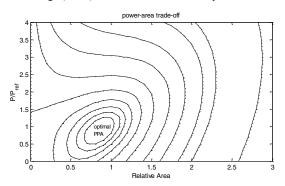

Figure 3: Normalized contours showing optimal power per area (PPA).

Design	Area		PPA		P_{total}/channel[µW]		SNR (100Hz-10kHz) [dB]/channel	
	WCD [mm²]	QPO rel.	WCD	QPO rel.	WCD slow, nom, fast [µW]	QPO rel.	WCD slow, nom, fast [dB]	QPO rel.
LNA	0.096	0.86	1	0.86	7.12, 7.15, 7.16	0.81	57.44, 59.65, 61.22	1.18
LPF	0.052	0.78	1	0.82	8.64, 8.84, 8.94	0.74	56.23, 57.76, 58.44	1.21
HPF	0.066	0.85	1	0.84	5.47, 5.65, 5.71	0.82	55.86, 57.69, 58.55	1.19
PGA	0.058	0.91	1	0.92	9.56, 9.76, 9.82	0.79	58.54, 59.34, 60.26	1.23
SAR_{comp}	0.036	0.86	1	0.91	3.14, 3.21, 3.24	0.83	55.46, 57.52, 58.21	1.24
SAR_{DAC}	0.074	0.92	1	0.96	3.56, 3.69, 3.72	0.87	57.21, 59.67, 60.93	1.19
SAR_{logic}	0.042	0.81	1	0.87	4.52, 4.56, 4.57	0.81	61.94, 63.21, 64.32	1.25
Total	0.424	0.76	1	0.81	42.01, 42.86, 43.16	0.82	54.76, 56.21, 57.48	1.16
Average (relative)		0.84	1	0.87		0.81		

TABLE I– SUMMARY OF THE ALGORITHM PERFORMANCE WITH 99% YIELD

The limit on power dissipated can be expressed as $(8kT)\times f(\text{SNR})$, where kT is the thermal energy, and f is an increasing function of SNR [24]. Additionally, the interface input to the neural system is subject to external noise, which can be represented by an effective temperature. Reducing noise to improve signal processing requires larger numbers of receptors, channels, or neurons, requiring additional power resources [25].

IV. CONCLUSION

In this paper, we develop a yield constrained sequential power per area minimization framework that is applied to a multivariable optimization in a neural recording interface. By limiting over-dimensioning of the circuit, the proposed method achieves consistently a better power per area ratio over the entire range of neural recording interface circuits, with no loss of circuit performance. Our approach can be used with any variability model and is not restricted to any particular performance constraint. As the experimental results in CMOS 90nm technology indicate, the suggested numerical methods provide accurate and efficient solutions of the power per area optimization problem offering up to 26% power savings and up to 22% area reduction, without yield penalties.

REFERENCES

[1] M.A. Lebedev, M.A.L. Nicolelis, "Brain-machine interfaces: Past, present and future," *Trends Neurosci.*, vol. 29, no. 9, pp. 536-546, 2006.

[2] G. Buzsaki, "Large-scale recording of neuronal ensembles," *Nat Neurosci*, vol. 7, pp. 446-451, 2004.

[3] F.A. Mussa-Ivaldi, L.E. Miller, "Brain-machine interfaces: Computational demands and clinical needs meet basic neuroscience," *Trends Neurosci.*, vol. 26, no. 6, pp. 329-334, 2003.

[4] M. Mollazadeh, K. Murari, G. Cauwenberghs, N. Thakor, "Micropower CMOS-integrated low-noise amplification, filtering, and digitization of multimodal neuropotentials," *IEEE Trans. Biomed. Circuits Syst.*, vol. 3, no. 1, pp. 1-10, 2009.

[5] A.M. Sodagar, *et al.*, "An implantable 64-channel wireless microsystem for single-unit neural recording," *IEEE J. Solid-State Circ.*, vol. 44, no. 9, pp. 2591-2604, 2009.

[6] B.K. Thurgood, *et al.*, "A wireless integrated circuit for 100-channel charge-balanced neural stimulation," *IEEE Trans. Biomed. Circuits Syst.*, vol. 3, no. 6, pp. 405-414, 2009.

[7] S. Kim, R. Normann, R. Harrison, F. Solzbacher, "Preliminary study of the thermal impact of a microelectrode array implanted in the brain," *IEEE Int. Conf. Engin.in Med. and Biol. Soc.*,pp. 2986-2989, 2006.

[8] X. Zou, *et al.*, "A 100-channel 1-mW implantable neural recording IC," *IEEE Trans. Circuits Syst.-I: Reg. Papers*, vol. 60, no. 10, pp. 2584-2596, 2013.

[9] C. Chae, *et al.*, "A 128-channel 6 mw wireless neural recording IC with spike feature extraction and UWB transmitter," *IEEE Trans. Neural Syst. Rehab. Engin.*, vol. 17, no. 4, pp. 312-321, 2009.

[10] J. Lee, H.-G. Rhew, D.R. Kipke, M.P. Flynn, "A 64 channel programmable closed-loop neurostimulator with 8 channel neural amplifier and logarithmic ADC," *IEEE J. Solid-State Circ.*, vol. 45, no. 9, pp. 1935-1945, 2010.

[11] R. Brodersen *et al.*, "Methods for true power minimization," *IEEE Int. Conf. Comp.-Aided Design*, pp. 35-42, 2002.

[12] A. Bhavnagarwala, B. Austin, K. Bowman, J.D. Meindl, "A minimum total power methodology for projecting limits on CMOS GSI," *IEEE Trans. VLSI Syst.*, vol. 8, no. 6, pp. 235-251, 2000.

[13] G. Yu, P. Li, "Yield-aware hierarchical optimization of large analog integrated circuits," *IEEE Int. Conf. Comp.-Aided Design*, pp. 79-84, 2008.

[14] F. Schenkel, *et al.*, "Mismatch analysis and direct yield optimization by specwise linearization and feasibility-guided search," *IEEE Design Autom. Conf.*, pp. 858-863, 2001.

[15] T. Mukherjee, L. R. Carley, and R. A. Rutenbar, "Efficient handling of operating range and manufacturing line variations in analog cell synthesis," *IEEE Trans. Comp.-Aided Design*, vol. 19, no. 8, pp. 825-839, 2000.

[16] S. Seth, B. Murmann, "Design and optimization of continuous-time filters using geometric programming", *IEEE Int. Symp. Circ. Syst.*, pp. 2089-2092, 2014.

[17] B. Gosselin, "Recent advances in neural recording microsystems," *Sensors*, vol. 11, no. 5, pp. 4572-4597, 2011.

[18] R.R. Harrison, *et al.*, "A low-power integrated circuit for a wireless 100-electrode neural recording system," *IEEE J. Solid-State Circ.*, vol. 42, no. 1, pp. 123-133, 2007.

[19] R.R. Harrison, G. Santhanam, K.V. Shenoy, "Local field potential measurement with low-power analog integrated circuit," *IEEE Int. Conf. Eng. Med. Biology Soc.*, vol. 2, pp. 4067-4070, 2004.

[20] A. Zjajo, *et al.*, "Stochastic analysis of deep-submicrometer CMOS process for reliable circuits designs," *IEEE Trans. Circuits Syst.-I: Reg. Papers*, vol. 58, no. 1, pp. 164- 175, 2011.

[21] Y. Freund, R.E. Schapire, "Large margin classification using the perceptron algorithm," *Mach. Learning*, vol. 37, pp. 277-296, 1999.

[22] I. Tsochantaridis, T. Hofmann, T. Joachims, Y. Altun, "Support vector machine learning for interdependent and structured output spaces," *Int. Conf. Machine Learning*, pp. 1-8, 2004.

[23] A. Dharchoudbury, S.M. Kang, "Worst-case analysis and optimization of VLSI circuits performances", *IEEE Trans. Comp.-Aided Design*, vol. 14, no. 4, pp. 481-492, 1995.

[24] E.A. Vittoz, "Future of analog in the VLSI environment", *IEEE Int. Symp. Circ. Syst.*, pp. 1372-1375, 1990.

[25] J.E. Niven, S.B. Laughlin, "Energy limitation as a selective pressure on the evolution of sensory systems," *J. Exp. Biol.*, vol. 211, no. 11, pp. 1792-1804, 2008.