Dynamic Power Reduction in Scalable Neural Recording Interface Using Spatiotemporal Correlation and Temporal Sparsity of Neural Signals

Sung-Yun Park, Jihyun Cho, Kyuseok Lee, and Euisik Yoon, *Member, IEEE*

***Abstract*— We report a scalable neural recording interface with embedded lossless compression to reduce dynamic power consumption (P_D) for data transmission in high-density neural recording systems. We investigated the characteristics of neural signals and implemented effective lossless compression for local field potential (LFP) and extracellular action potential (EAP or spike) in separate signal paths. For LFP, spatial–temporal (spatiotemporal) correlation of the LFP signals is exploited in a Δ-modulated ΔΣ analog-to-digital converter (Δ − ΔΣ ADC) and a dedicated digital difference circuit. Then, statistical redundancy is further eliminated through entropy encoding without information loss. For spikes, only essential parts of waveforms in the spikes are extracted from the raw data by using spike detectors and reconfigurable analog memories. The prototype chip was fabricated using 180-nm CMOS processes, incorporating 128 channels into a modular architecture that is easily scalable and expandable for high-density neural recordings. The fabricated chip achieved the data rate reduction for the LFPs and spikes by a factor of 5.35 and 10.54, respectively, from the proposed compression scheme. Consequently, P_D was reduced by 89%, when compared to the uncompressed case. We also achieved the state-of-the-art recording performance of 3.37 μW per channel, 5.18 μV_{rms} noise, and 3.41 $\text{NEF}^2 V_{\text{DD}}$.**

***Index Terms*— Dynamic power, extracellular action potential (EAP or spike), high-density neural recording, local field potential (LFP), lossless compression, modular, neuroscience, scalable, spatiotemporal correlation, temporal sparsity.**

I. Introduction

OVER the past years, neural recording interface systems have significantly progressed to provide high-quality, parallel recordings of neural activities at low power in a small form factor, while dramatically increasing the number of channels. According to a recent survey, the number of parallel recording channels has steadily increased, doubling every seven years similar to Moore's law although much slower, since the initial demonstration of the unit recording in 1960 [1]. The current state-of-the-art neural recording interface systems have acquired a capability to facilitate a few hundred of parallel recordings [2]–[4], and it is projected to reach to >1000-channel parallel recording in a few years. This advancement is mainly stimulated by the fact that large-scale recording of neuronal ensembles can provide the in-depth understanding of brain activity [5], [6]. In order to meet this aggressive scaling, neural recording interfaces should be able to provide high-quality recordings which can fully embrace the broadband waveforms, both local field potentials (LFPs) and APs (action potentials or spikes) with limited energy and area budget. Utilizing various circuit techniques, the performance of core parts in neural recording interfaces [analog front-end (AFE)] has steadily been improved [2]–[4], [7]–[15]. For instance, a recent neural interface achieved sub-microwatt per channel of power consumption in the AFE with reasonable performance [12] and others demonstrated the implementation of AFE in an extremely small area of 0.013 mm^2 per channel [13], [14]. The previous efforts showed the scalability of neural interface circuits in either power or area as a promising milestone toward massively parallel recording systems.

As opposed to the efforts for developing of energy- and area-efficient AFEs, the associated digital circuits for processing and transferring the recorded data in neural recording interfaces have been less focused. Obviously, as the number of simultaneous recording channels increases, the data rate of the recorded signals also keeps escalating. The increased data transmission rate, i.e., larger data bandwidth inevitably leads to the higher power consumption of digital circuits in handling and transmitting the huge amount of data. In fact, it is highly possible that the power consumption of digital circuits becomes much larger than that of the AFEs. By a simple calculation based on several assumptions, we can easily estimate this trend in scaling, as shown in Fig. 1. A 128-channel parallel recording interface system generates a serialized data rate of 32 Mb/s if we assumed 10-b data streams in each channel with a sampling rate of 25 kS/s. For data transmission only, it may consume ~4 mW when using a standard CMOS I/O buffer (with 15 pF load capacitance). The actual power consumption in the digital circuits that support AFEs is even higher because it needs other data processing and

Manuscript received August 22, 2017; revised October 25, 2017 and November 25, 2017; accepted December 15, 2017. Date of publication January 23, 2018; date of current version March 23, 2018. This paper was approved by Guest Editor Makoto Ikeda. This work was supported by NSF 1545858. *(Corresponding author: Euisik Yoon.)*

S.-Y. Park, K. Lee, and E. Yoon are with the Center for Wireless Integrated MicroSensing and Systems, Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: sungyun@umich.edu; eekslee@umich.edu; esyoon@umich.edu).

J. Cho is with Apple Inc., Cupertino, CA 95014 USA (e-mail: jihyun_cho@apple.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2017.2787749

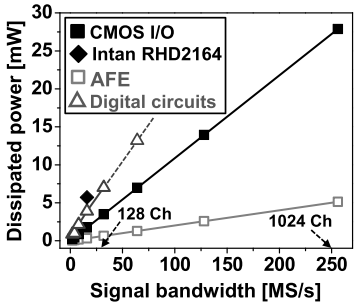


Fig. 1. Power consumption versus signal bandwidth (number of parallel recording). A commercial product (Intan RHD2164) is also included.

control units. In our previous prototype [11], it was measured about ~7 mW. When compared with the current state-of-the-art AFE's power consumption (assumed 5 μW/Ch for Fig. 1), it is more than 10 times larger. If the ultimate goal is to build a neural recording interface system having >1000 simultaneous recording channels, the power consumption of the digital circuit should be significantly reduced. Otherwise, it will become a major bottleneck of the scaling in total power consumption.

In this paper, we propose an embedded lossless data compression scheme for both LFPs and spikes to reduce the dynamic power consumption of digital circuits (P_D) in neural interface systems [16]. The proposed scheme is efficient and effective in terms of utilizing system resources such as area and power consumption while providing high-quality recordings. In addition, this scheme can be applied to either wireless or wireline data transmission in neural recording systems because the compression scheme we developed is generic and does not depend on specific data transmission methods. For the compression, we exploited the inherent characteristics of neural signals: spatiotemporal correlation of LFPs and temporal sparsity of spikes. In this paper, we implemented a 128-channel neural recording chip for proof of concept, but this interface system is easily scalable to a higher number of channels (1024 channels and beyond) since each recording and compression block is modular and expandable (independently operating within a subgroup of channels).

This paper is organized as follows. The inherent characteristics of LFPs and spikes are investigated and the optimal scheme for lossless compression is discussed in Section II. In Section III, the architecture for a scalable 128-channel neural recording interface with embedded lossless compression is explained. Section IV provides the detailed description of essential integrated circuit blocks used to realize the proposed recording and compression scheme. The measurement results obtained from *ex-vivo* electrical characterization and *in vivo* animal experiments are presented in Section V. The summary and comparison of the measured performance with other state-of-the-art works are also given in Section V. Finally, Section VI concludes this paper.

II. Investigation of Broadband Neural Signals

To understand in-depth brain activity, both LFPs and spikes should be simultaneously recorded [5]. The sum of the two signals constitutes a wide bandwidth from sub-hertz (0.1–0.5 Hz) to a few kilohertz (5–7.5 kHz) as well as a wide dynamic range from a few tens of microvolt to millivolt (100 μV–3 mV, typically). The wideband, wide dynamic range neural signals impose stringent requirements in recording system implementation: >10-b resolution with <10-$\mu V_{\rm rms}$ noise while consuming a few microwatt power. Recent works have tried to alleviate those requirements by utilizing the inherent characteristics of neural signals, $1/f$ power spectra, for processing broadband neural signals [11], [17] and Electrocorticography (ECoG) [18]. In this paper, we investigate other characteristics of neural signals to explore the possibility to implement lossless compression in addition to the $1/f$ feature.

A. Spatiotemporal Correlation in Local Field Potential

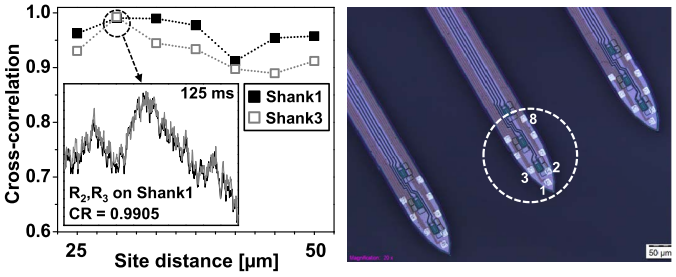
LFPs are the ensemble collection of neuronal activities from multiple nearby neurons. Signals have inherent $1/f^n$ $(n = 1 - 2)$ power spectra [6]. This $1/f$ spectrum indicates that the LFP signals have higher energy in their low-frequency range, i.e., LFPs have a high temporal correlation. According to the previous works, the overall shape of broadband (LFP + spikes) neural signal power spectra looks roughly $1/f$, but that is because of the high-energy contents of the LFPs in the broadband signal. Therefore, we can naturally guess that if the LFPs are isolated by proper filtering, they have a higher temporal correlation. Then, by taking the temporal difference of LFPs, effective data compression can be achieved.

Also, we can surmise another inherent characteristic of LFPs. Slow-varying LFPs are the aggregated sum of the signals from local neurons. Since they are the so-called "average values" from multiple neurons in proximity, the spatial correlation between the local LFPs from the neighboring recording sites must be high. In particular, since the distance between the recording sites becomes smaller to record the activities of more neurons in a given volume in advanced neural probes [19], the spatial correlation of LFPs recorded from neighboring channels can become higher. As most lossless compression schemes use the statistical redundancy to represent data, this high spatial correlation in the LFPs also suggests that we can compress LFPs losslessly using this feature. To verify this guess, we calculated spatial correlations of LFPs. Fig. 2(a) and (b) shows the spatial correlation of the adjacent LFPs from two sets of the eight recording sites (eight recording sites on each shank) and a microphotograph of the probe used to record the LFPs, respectively. The signals were pre-recorded from the CA1 region of a mouse brain for 2 s, and low-pass filtered by a commercial software (MATLAB). The distance between the recording sites linearly increases from 25 to 50 μm. As shown, the spatial correlation of the LFPs from the adjacent recording sites is mostly >0.9.

By using the spatial and temporal correlation of the recorded LFPs, we calculated the compression rate (CR) of LFPs, which is given by

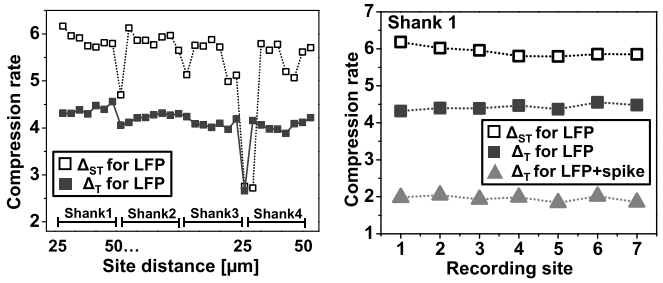
$$\mathrm{CR} = \frac{N_o}{N_e} \tag{1}$$

where N_o and N_e are the numbers of bits to represent the original and encoded data sets, respectively. After filtering,



(a) (b)

Fig. 2. (a) Spatial correlation of LFPs from the nearby recording sites. (b) 32-channel electrical probe used for recording of the neural signals.



(a) (b)

Fig. 3. (a) Compression rate for temporal or spatial difference of LFPs. (b) Spatiotemporal and temporal difference of LFPs, and temporal difference of broadband neural signals.

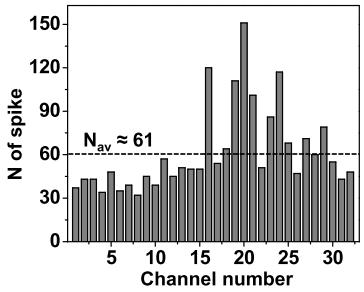


Fig. 4. Number of spikes recoded from the 32-channel, eight-shank neural probe.

we took two differenced signals of LFPs: one is the temporal differenced (Δ_T) and the other is spatiotemporal differenced ($\Delta_{ST} = \Delta_S + \Delta_T$) LFPs. The differenced signals were quantized into 10 b and applied to an ideal entropy encoder (Huffman encoder in this case). Fig. 3(a) shows the calculated *CRs* of LFPs from the 32 recording sites distributed across four different shanks [see Fig. 2(b)]. The duration of LFPs is 2 s. The average *CR*s when Δ_T and Δ_{ST} are applied were about 4.3 and 5.8, respectively. As expected, the spatiotemporal correlation offers a higher *CR*. In fact, the *CR* could have been much higher if we had taken spatiotemporal differences within a shank since the spatial correlation between the sites in different shanks may not be high due to a longer site distance. In this calculation, we included the shank-to-shank differences, shown as a sudden drop in Fig. 3(a), to comprehensively explain that the compression is affected by the spatial correlation, in other words, the site distance. Moreover, we included the worst case of a non-working site (the channel #24) that shows the big drops in Fig. 3(a).

We also evaluated the *CRs* comparing two cases: 1) LFPs only and 2) broadband signals (LFPs+spikes). This will verify whether the LFPs have a higher temporal correlation than the combination of LFPs and spikes or not. As shown in Fig. 3(b), the *CRs* of Δ_T for LFPs only ($\sim$4.43) are more than two times higher than those of the broadband signals ($\sim$1.95). This can be explained by the fact that spikes are single cell activities and the firing rate is sparse and rather uncorrelated to one another and also to LFPs [6]. Based on our numerical calculation and validation, we can conclude that it is desirable to separate LFPs from spikes for better compression.

Various data compression schemes, either lossy or lossless, have been investigated to reduce the data rate for bio-signals [20]–[22]. Depending on applications, lossy compression is preferred. In [23], the compression was achieved with discrete packet wavelet transforms. The modeling with antireflection all-pole filters was also implemented while assuming the electromyography signals are wide-sense-stationary [24]. For lossless compression, the electroencephalography (EEG) signals are compressed using spatiotemporal correlations [25]. However, there has relatively been little research on LFP compression, although signal properties are similar to EEG. The validity of our approach for lossless compression of LFPs is rooted in the basic principles in EEG compression [25], either spatial or temporal correlation, or both.

B. Temporal Sparsity of Spike

Spikes are known to be sparse in the time domain, and their firing rate is activity dependent. Spikes have a short waveform for a few milliseconds of duration (3–4 ms, typically), and their rate of occurrence ranges from a few tens to hundreds per second [26], [27]. Although there are many studies about the sparsity of spikes and the applications using this characteristic, we conducted our own numerical calculation to evaluate the sparsity of spikes for effective compression of the spike signals. Fig. 4 shows the number of spikes from the 32-channel probe for 2 s. The spikes were extracted from the broadband neural signals using a software high-pass filter and the amplitude thresholding with a median value of background signals [or inter-spike intervals (ISI), mostly noise] [27]. The average firing rate of the spikes was found as only 61, very sparse compared to the entire 40 000 points in the data set (equivalently, $\sim$3% of duty).

Some of the recent neural interfaces for spike recording and compression have exploited this sparseness to reduce the data rate [28]–[30]. In those, the filtering of broadband neural signals is performed first, then the detection and separation of spikes from the ISIs are achieved through the analog, digital, or mixed-signal spike detectors. After the spikes are separated from the background noise, only the timing information of spikes, ISIs, or both are sent to the off-chip host system to minimize the data transmission rate [31]. However, this approach inherently accompanies the loss of information, such as the shapes or derivatives of the spikes, which are useful clues for follow-up neuroscience studies. To preserve the entire waveforms of the spikes, the compression scheme should employ the dedicated memories in the analog or

digital domain. Otherwise, the part of spike waveforms (called as the preambles of spikes) can be lost due to the detection latency. In order to preserve the integrity of spikes, the length of preambles of spikes should be at least 500–800 μs [27], equivalent to about 12–18 samples assuming that the sampling rate is 20 to 31.25 kS/s. In [29], the dedicated 16-B memory per channel (16 data storage spaces with 8-b resolution) was implemented using the static random access memory (SRAM) and demonstrate the feasibility in *in vivo* environments. However, the SRAM in the implementation is bulky and the analog-to-digital converter (ADC) should constantly operate and consume power even though there is no spike. The other way proposed to capture the entire waveforms of spikes is the two-signal paths approach: fast and slow paths [30], [32]. The fast path was for detection of spikes, and the slow path was used for propagation of the replica of spikes with ~600 μs of analog delay. Nonetheless, the delayed waveform was distorted due to the ripples in the implementation, deteriorating the signal integrity. Recently, analog memories using on-chip capacitors are reintroduced to preserve the entire waveform of spikes [28]. However, the size of the on-chip capacitors used for the memory is too big and design was not optimized. Our implementation in this paper preserves signal integrity while using the small size of on-chip capacitors and also provides the reconfigurability to provide more flexibility for users.

III. Proposed 128-Channel Recording Interface With Embedded Compression

A scalable 128-channel neural recording system with embedded compression scheme has been designed to reduce the dynamic power consumption of the digital circuits for data handling and transmission while minimizing the additional system resources. For LFPs, we achieve a spatiotemporal compression in three consecutive steps by: 1) taking temporal differences using $\Delta - \Delta\Sigma$ ADC [11]; 2) taking spatial differences using digital circuits; and 3) using entropy encoders. For spikes, we provide the two modes of operation (compression and normal) by reconfiguring an in-channel successive approximation register (SAR) ADC. In the compression mode, the waveforms of spikes are extracted from the analog first-in-first-out (A-FIFO) memory only when spikes are detected. The spike detection is implemented by a simple thresholding method [27]. This thresholding scheme has an advantage that it does not miss spikes, but inevitably this may result in false positive and increase power consumption and bandwidth. However, the overhead is acceptable and much less than implementing a sophisticated spike detection algorithm on-chip. The A-FIFO memory reuses the capacitor digital-to-analog converter (CDAC) dedicated in the in-channel SAR ADC. On the other hand, in the normal mode (no-compression), the full raw spikes including ISIs are digitized in the SAR ADC and then sent to the off-chip host. Even though ISIs are conventionally regarded as noise, the no-compression mode is prepared for the case where the unexpected demand exists because the ISIs are permanently lost in the compression mode.

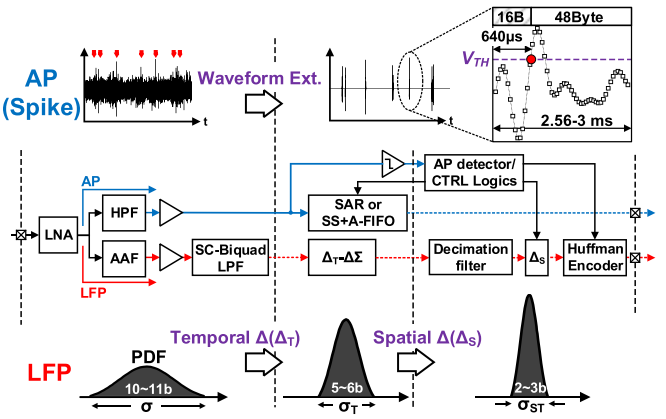


Fig. 5. Conceptual diagram for the compression in the proposed recording interface.

Fig. 5 conceptually visualizes the signal paths and lossless compression in the proposed neural interface. Once the neural signals including LFPs and spikes are pre-amplified in the Low-noise amplifier (LNA), each of those takes an independent processing path as indicated in Fig. 5. For LFPs, additional filtering and amplification are performed through a switched-capacitor low-pass filter (SC-LPF), followed by the time difference (Δ_T) and quantization ($\Delta\Sigma$) in the $\Delta_T - \Delta\Sigma$ ADC. The output of the ADC is filtered again in a decimation filter to reduce the data rate; then, the spatial differences (Δ_S) between channels (channel-to-channel difference) are taken in the digital circuit. Finally, the Huffam encoder removes the redundancy of the codes. The conceptual probability density functions during the LFP compression, shown in Fig. 5, illustrate how the LFP signals are compressed. The resolution of the raw LFPs is known as 10–11 b. Once Δ_T is taken, the distribution of the LFP becomes narrow because the amplitudes of time-differenced slow-varying signals decrease. The estimated number of bits resulted from the Δ_T compression is ~5–6 b. After that, the LFP signals are further compressed in the digital domain through the Δ_S compression, by taking advantage of spatial correlation among the recording sites. The signals are compressed down to ~2–3 b after removing redundancy by the Huffman encoder. For spikes, we use a continuous time-high-pass filter to isolate them from LFPs. The filtered and amplified raw spikes (spikes + ISIs) are saved in the analog memory inside the single slope (SS)/A-FIFO block without digitization unless there is an onset signal in the spike detection block (a comparator). Once a spike detected, the SS-ADC retroactively converts the data in the memory for the certain amount of time (3–10 ms) and send them to the off-chip host system. Since the raw data is stored in the A-FIFO memory, the preambles of spikes are preserved, achieving lossless compression.

Fig. 6 shows the top-level architecture and circuit block diagram of a scalable 128-channel neural recording interface with the embedded compression scheme. Each 16-channel forms a single group in the interface. Each group operates independently so that this modular design is easily scalable to any multiples of 16. A single-channel AFE consists of an LNA, a low- and high-pass filters (LPF/HPF) with programmable gain and bandwidth, and a spike waveform

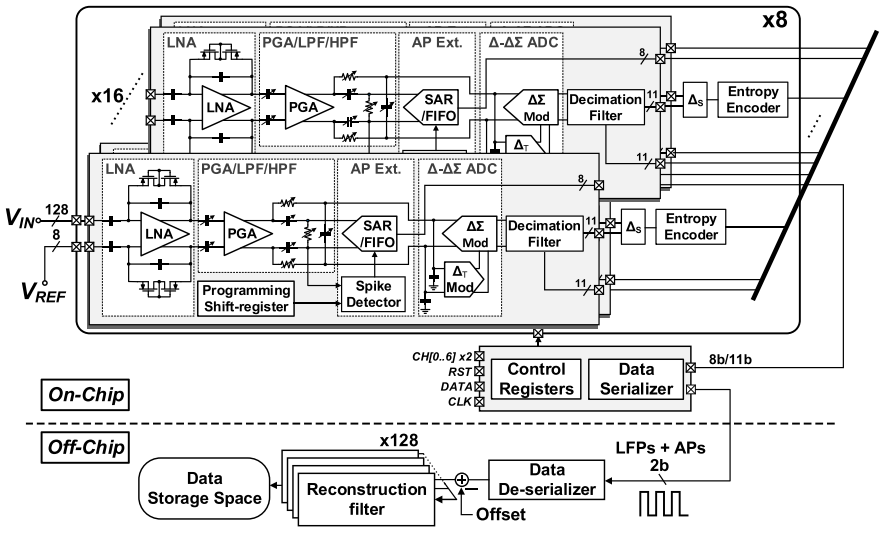


Fig. 6. Top-level circuit architecture of a 128-channel neural recording interface with embedded lossless compression scheme.

extractor which consists of a reconfigurable SAR or SS-ADC+ A-FIFO, a $\Delta-\Delta\Sigma$ ADC, and a digital decimation filter. All analog and mixed signals are processed in a fully differential manner to maximize the signal dynamic range. There are bias circuits in every 16 channels, and a spike detector resides in each spike waveform extractor. There are also programming shift registers to change the gain and bandwidth settings of LNA, programmable gain amplifier (PGA), the threshold voltage of the spike detector, and the length of spikes. All of the processed LFPs and spikes are serialized onto two wires. The off-chip signal restoration system is also included in Fig. 6 to illustrate the signal reconstruction. The reconstruction process is necessary only for the compressed LFPs.

IV. Circuit Block Implementation

In this section, the details of the circuit implementations for individual blocks are explained. Some trivial circuits such as an anti-aliasing filter (AAF) and buffers are not included.

A. Low-Noise Amplifier

Since the LNA is the first block of the whole signal processing chain, it must provide high gain and low noise performance to preserve the signal integrity of an entire recording channel. In addition, because the LNA is interfacing with the electrodes, it should reject large dc fluctuations from the electrode–electrolyte interface [26]. We employed a capacitive-coupled closed-loop feedback [33] and a two-stage Miller-compensated operational transconductance amplifier (OTA) for the LNA, as shown in Fig. 7. $V_{\rm DD}$ is the power supply of LNA and $V_{\rm CM}$ is the common mode voltage, which is a half of $V_{\rm DD}$ in Fig. 7. $V_{\rm CM0}$ ($\sim$0.21 V, programmable) is intended for the bias of the input PMOS transistors ($M_{3,4}$) because $V_{\rm CM}$ cannot provide a proper input bias voltage for the PMOS transistors. The transfer function of the LNA is given by [34]

$$\frac{V_{\rm out}}{V_{\rm in}} \approx \frac{C_{\rm in}}{C_{\rm fb}} \frac{sC_{\rm fb}R_{\rm pseudo}\left(1-s\frac{C_{\rm fb}}{G_{m1}G_{m2}r_{\rm out1}}\right)}{(1+sC_{\rm fb}R_{\rm pseudo})\left(1+s\frac{C_{\rm in}}{C_{\rm fb}}\cdot\frac{C_C}{G_{m1}}\right)} \tag{2}$$

where $C_{\rm in}$, $C_{\rm fb}$, C_C, and $R_{\rm pseudo}$ are the input, feedback, compensation capacitors, and feedback resistor, respectively, and G_{m1} and G_{m2} are the transconductances of the 1st- and 2nd stages of the OTA used for the LNA. A sub-Hz high-pass corner frequency (f_L) to suppress dc fluctuations is formed by $C_{\rm fb}$ and $R_{\rm pseudo}$, which is designed as 0.4 Hz. The mid-band gain of the LNA is solely set by the closed-loop feedback formed with $C_{\rm in}$ and $C_{\rm fb}$ if the gain of the OTA is large enough. The value of $C_{\rm in}$ was selected as 5.85 pF by considering design specifications such as gain, noise, and input impedance [35]. The transistor level implementation of the two-stage OTA used for the LNA is also shown in Fig. 7 (right). The dc gain of the OTA is $\sim$70 dB in the simulation. The total input referred noise (IRN) density of the OTA is approximately given by

$$\overline{v^2_{\rm ni,OTA,theraml}} \approx \frac{8kT}{3}\cdot\left(\frac{1}{g_{m1}+g_{m3}}\right) \tag{3a}$$

$$\overline{v^2_{\rm ni,OTA,1/f}} \approx \frac{1}{C_{\rm ox}}\cdot\left(\frac{K_n g^2_{m1}}{(WL)_1}+\frac{K_p g^2_{m3}}{(WL)_3}\right)\cdot\left(\frac{1}{g_{m1}+g_{m3}}\right)^2\cdot\frac{1}{f} \tag{3b}$$

where k is Boltzmann's constant, T is the absolute temperature in Kelvin, $C_{\rm ox}$ is the oxide capacitance per unit area, K_n and K_p are the process-dependent $1/f$ noise parameters for NMOS and PMOS transistors for the given 180-nm CMOS process, and g_{m1} and g_{m3} are transconductances of M_1 and M_3, respectively. To achieve Low-noise performance, the input transistors (M_{1-4}) in the OTA are designed to operate in the deep subthreshold region, where transconductance efficiency (g_m/I_D) is maximized [36]. In addition, the two complementary inputs ($M_{1,3}$ and $M_{2,4}$) can increase the transconductance of the first stage by a factor of two, resulting in a reduction of input referred thermal noise voltage by a factor of $\sim\sqrt{2}$, as given by (3a). To reduce $1/f$ noise, the large area of the input transistors (800/0.35 μm and 840/0.25 μm for $M_{1,2}$, and $M_{3,4}$, respectively) are used. The low supply voltage

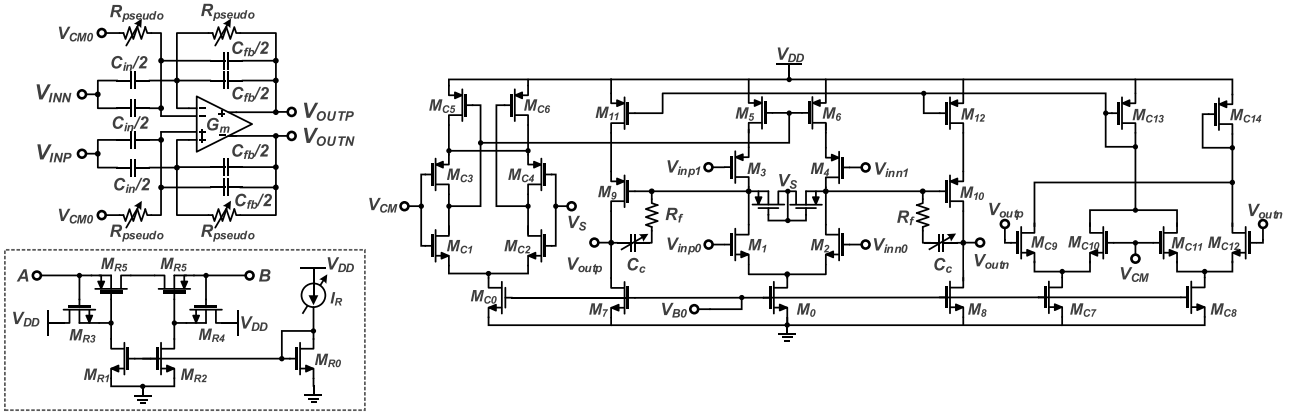


Fig. 7. Schematic of LNA (top left) and pseudo-resistor (bottom left) used for the LNA, and schematic of the OTA (right) used for the LNA.

may deteriorate the common mode rejection ratio (CMRR), so the dual tail currents (M_0, M_5, and M_6) are used to reduce the CM gain. We also implemented two common mode feedbacks (CMFBs) in each stage to prevent the latch-up during startup [35]. The CMFB circuits are shown in Fig. 7 by denoting the used transistors as Mc_0–Mc_6, Mc_7–Mc_{14}. The CMFB circuits are based on a conventional common-source single-stage amplifier (left) and a common-source differential-difference amplifier (right). Since R_{pseudo} is vulnerable to process variations, it is desirable to make it programmable. In lieu of using simple PMOS transistors, where the gate voltage is used for tuning [33], we adopted a pseudo resistor structure where source-to-gate voltage (V_{SG}) of $M_{R5,6}$ is controlled by current to achieve better linearity [37], as shown in Fig. 7 (bottom left). The LNA consumes $\sim$2.09 μA of current and the important performance metrics, such as NEF and NEF$^2 V_{\text{DD}}$, are comparable with other state-of-the-art works.

B. Low-and High-Pass Filter With Programmable Gain and Bandwidth

The LFPs and spikes are separated after initial amplification, and different compression schemes are applied, respectively. As briefly mentioned in Section III, the LFPs are isolated using the SC-biquad LPF with a programmable gain and bandwidth after rejecting high-frequency signals by using the AAF. The target corner frequency (f_{LPF}) to extract LFPs from the broadband neural signals is known as 150–300 Hz [26]. To realize such a low frequency with a CT-filter, bulky passive components are usually required. Since one of our primary goals in the proposed design is to minimize the system resources, the filter was implemented using SC design techniques to reduce area consumption. Even though the bandwidth requirement in the OTA for SC-filters is larger than the one for CT-filters to secure stable settling within the given sampling periods, the increased power consumption affects little to the overall, as f_{LPF} is only a few hundred Hz (13.1% of the total AFE power consumed in the SC-LPF). Fig. 8(a) shows the schematic of the implemented SC-biquad LPF. The frequency of the non-overlapping clocks (f_{CLK}), φ_1 and φ_2, was set as 64 kHz, considering a sampling frequency of 2 kHz and an oversampling ratio (OSR)

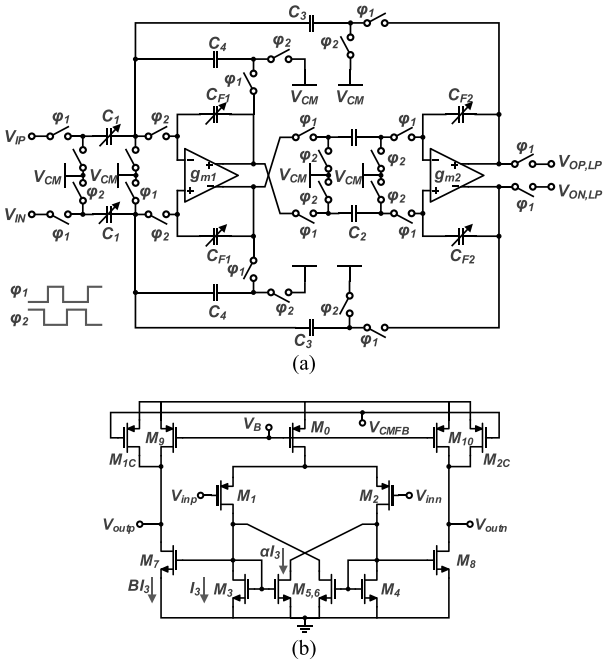


Fig. 8. (a) Schematics of the SC-biqud LPF and (b) OTA for the LPF.

of 32. The SC-LPF has four programmable gains (A_{LPF}): 0, 6, 9.5, and 12 dB, and four f_{LPF} options: 150, 200, 250, and 300 Hz. Users can adjust the ratio of C_1 to C_3 and the feedback capacitors, C_{F1} and C_{F2}. The A_{LPF} and f_{LPF} of the SC-biquad LPF are determined as

$$A_{\text{LPF}} = \frac{C_1}{C_3}, \quad f_{\text{LPF}} = \frac{1}{2\pi}\sqrt{\frac{C_2 \cdot C_3}{C_{F1} \cdot C_{F2}}} f_{\text{CLK}} \tag{4}$$

where C_1 changes from 40.8 to 163.2 fF for the variable gain and C_3 is 40.8 fF. The quality factor (Q_{LPF}) of the filter is also given by

$$Q_{\text{LPF}} = \sqrt{\frac{C_{F1}}{C_{F2}}} \cdot \sqrt{\frac{C_2 \cdot C_3}{C_4^2}} \tag{5}$$

where C_4 is set as 56.4 fF. According to (4) and (5), f_{LPF} can be adjusted by keeping Q_{LPF} (constant ratio of C_{F1} to C_{F2})

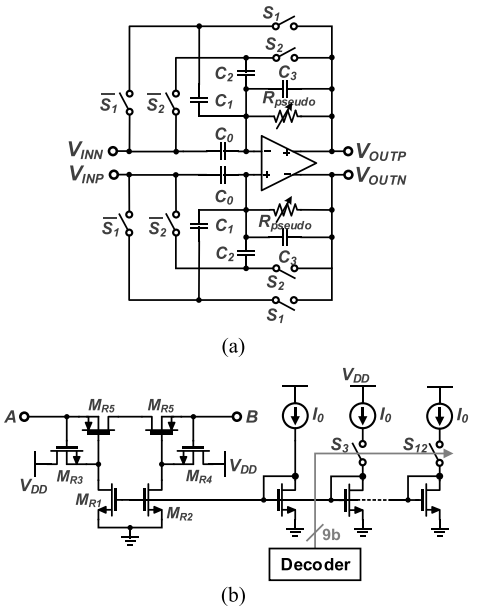


Fig. 9. (a) Schematics of the HPF and (b) pseudo-resistors used for the tuning of f_{HPF}.

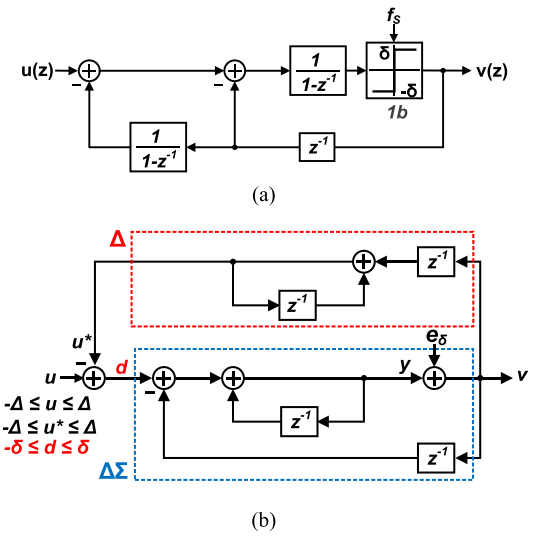


Fig. 10. (a) z-domain block diagram of the DT $\Delta-\Delta\Sigma$ ADC. (b) Expanded z-domain block diagram of the DT $\Delta-\Delta\Sigma$ ADC.

as a constant. In this design, Q_{LPF} is set as 0.7 for the flat magnitude response. Fig. 8(b) shows the schematic of the OTA used for the SC-LPF. To enhance the gain and bandwidth of the OTA, the local positive feedback ($M_{5,6}$) is used for the OTA [38]. The gain (A_O) and gain-bandwidth product (GBW) are given by

$$A_O = g_{m1} R_{\mathrm{out}} \cdot \frac{B}{1-\alpha}, \quad \mathrm{GBW} = \frac{g_{m1}}{2\pi C_L} \cdot \frac{B}{1-\alpha} \tag{6}$$

where α and B are the ratio of the transistors, $M_{5,6}$ to $M_{3,4}$ and $M_{3,4}$ to $M_{7,8}$, respectively, g_{m1} is the transconductance of M_1 (or M_2), and R_{out} is the output impedance of the OTA. In this design, $\alpha = 3/5$ and $B = 8/5$. The power consumption of the SC-biquad LPF is $\sim$0.444 μW.

The HPF with programmable gains and bandwidths for spike isolation is implemented by adopting a flip-over-capacitor topology [39], as shown in Fig. 9. It has four different gains: 7.5, 11, 17, and 23 dB, and nine different f_{HPF} settings between 80 and 600 Hz. Since the amplitude of spikes is relatively smaller than that of LFPs, we designed higher gains for the spike amplification in the HPF. The best cutoff frequency for spike isolation is debatable; however, the adjustable high-pass cutoff frequency f_{HPF} can fully embrace the necessary amount of signals up to 600 Hz. The gain is tuned by changing the ratio of the input to feedback capacitors ($S_{1,2}$ and their complements) as depicted in Fig. 9(a) and f_{HPF} is controlled by the variable pseudo-resistors via the current DAC (S_3–S_{12}) as shown in Fig. 9(b). The values of capacitors: C_0, C_1, C_2, and C_3 are 950, 55, 180, and 55 fF, respectively. The current mirror OTA is used for the HPF because of its simplicity. The power consumption of the HPF is $\sim$0.275 μW.

C. Discrete-Time $\Delta-\Delta\Sigma$ Analog-to-Digital Converter

For the quantization ($\Delta\Sigma$) and time difference (Δ_T) of LFPs, a discrete-time (DT) $\Delta_T-\Delta\Sigma$ ADC is employed [11], [34]. Fig. 10(a) shows the z-domain block diagram of the DT $\Delta-\Delta\Sigma$ ADC consisting of two integrators, each for Δ_T and $\Delta\Sigma$ operation, respectively, and a 1-b quantizer (a dynamic comparator). Fig. 10(b) is equivalent to Fig. 10(a), but it separates the Δ_T operation in order to better illustrate the operation of $\Delta-\Delta\Sigma$ ADC. The input and output signals of $\Delta_T-\Delta\Sigma$ are denoted by u and v, respectively, and the dotted lines indicate the Δ_T operation (red) and the 1st-order $\Delta\Sigma$ ADC (blue) in Fig. 10(b). Since the signal transfer function (STF) is unity in the $\Delta\Sigma$ ADC with only delay (STF $= z^{-1}$), the input of the $\Delta\Sigma$ ADC, d, appears at the output of the ADC, v. The input signal d is then integrated and fed back to the input; thus, it becomes a delayed replica of u, $u*$. Assuming that the range of u is confined within $|\Delta|$, the returning signal, $u*$, will also be in the same range as u. Because the $\Delta\Sigma$ ADC takes a difference (Δ_T) between u and $u*$, the time-differenced signal denoted as d has a smaller range of $|\delta|$ than $|\Delta|$, i.e., the compression can be achieved. After the d is quantized and sent to the off-chip system, the original signal u can be reconstructed by applying integration in the time domain (Σ_T), which is an inverse function of Δ_T. Even though the STF and noise transfer function (NTF) of the $\Delta_T-\Delta\Sigma$ ADC seem to be $1-z^{-1}$ and $(1-z^{-1})^2$, respectively, in this topology, the overall STF and NTF are 1(with delay) and $(1-z^{-1})$, respectively, if including the off-chip integration to restore the input signals.

As mentioned, the CT implementation is more energy-efficient for the oversampled ADCs; however, its component sizes become large if the CT implementation is adopted. In addition, the channel-to-channel variation can be reduced in the DT implementation since the matching property of the on-chip metal–insulator–metal (MIM) capacitors is better than that of the on-chip resistors [40]. Fig. 11 shows the schematic of the DT $\Delta_T-\Delta\Sigma$ ADC. The nominal values of capacitors, C_1, $C_{2,5}$, C_3, C_4, and C_F are set as 200, 100, 250, 50, and 1000 fF, respectively, and C_F is variable to cope with process variations. V_{REFP}, V_{REFN}, and V_{CM} are the reference and common-mode voltages, and the non-overlapping clock is

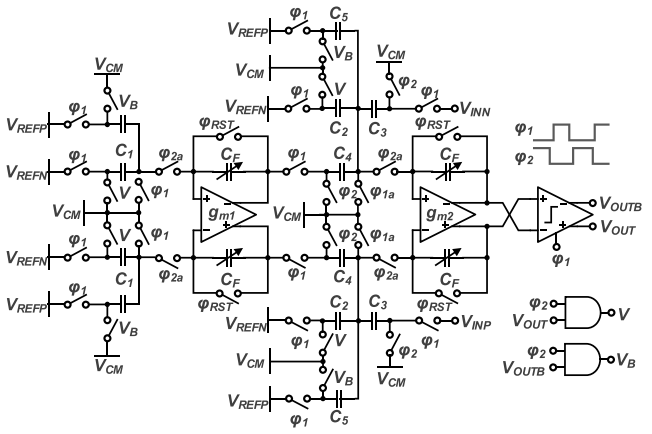


Fig. 11. Schematic of the DT $\Delta - \Delta\Sigma$ ADC.

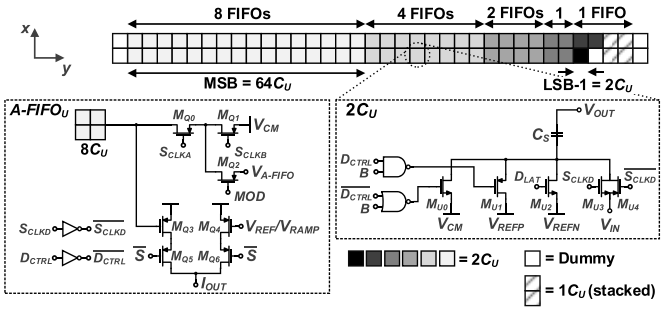


Fig. 12. A-FIFO (CDAC) configuration for the SS or SAR operation.

shared with the SC-LPF. The OTAs used in the DT $\Delta - \Delta\Sigma$ ADC denoted as g_{m1} and g_{m2} in Fig. 11. They have the same circuit topology as the one in Fig. 8(b) while having different design parameters such as gain, bandwidth, and noise.

D. Reconfigurable SS/SAR Analog-to-Digital Converter

There are two modes of operation: normal and compression modes for the recorded spikes as explained in Section II. In the normal mode, all the recorded spikes and ISIs are digitized by an in-channel 8-b SAR ADC without compression. On the other hand, in the compression mode, the incoming signals are sampled and temporally stored in A-FIFOs, and then converted into digital formats only when requested. As opposed to the recent literature where the dedicated analog (on-chip MIM capacitors) [28] and digital (SRAM) storages are implemented [29], we take an approach to have the A-FIFO reconfigured from the CDAC of the SAR ADC. Only when the spike detector issues the onset signals, the sampled data on the CDAC is quantized by comparing the data with a ramp signal (V_{RAMP}), which is an SS-ADC. The comparator used for this purpose also comes from the same one used in the SAR ADC. Fig. 12 shows the capacitor configuration and related transistors used for this purpose. The value of the unit capacitor for the A-FIFO$_U$ ($C_{U,\mathrm{FIFO}}$) must be determined by considering the leakage current, required resolution, and the duration of the preambles for spike detection. Since the acceptable length of the preamble is $\sim$640 μs [27], $C_{U,\mathrm{FIFO}}$ must preserve the stored data within the limit of the required

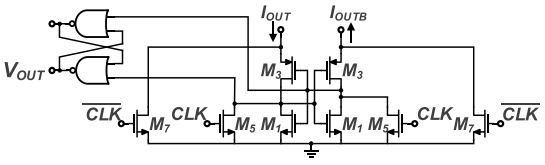


Fig. 13. Cross-coupled pair used for the dynamic comparator in SS/SAR ADC.

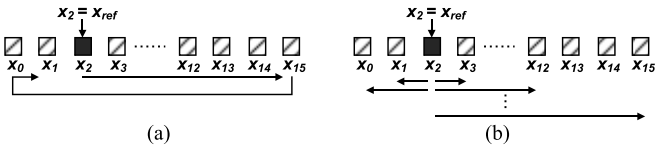


Fig. 14. Spatial difference of LFPs. (a) Center around and (b) adjacent channel difference.

8-b resolution during that time period. In the simulation, the data degradation is found to be less than a half of LSB (1 LSB $\approx$ 4 mV) during 640 μs if $C_{\mathrm{U,FIFO}}$ is larger than 180 fF. We set the value of $C_{\mathrm{U,FIFO}}$ as 182.4 fF by using MIM capacitors. In addition, because the total 16 different storage places (640 μs $\times$25 kS/s) are necessary at 25 kS/s sampling rate, total necessary capacitance is $\sim$2.92 pF and we can also set the unit capacitor (C_U) for the SAR ADC as 22.8 fF ($2.92/2^7$ pF). Fig. 12 highlights two circuits: $2C_U$ and unit A-FIFO (A-FIFO$_U$). The $2C_U$ has three switches (M_{U0-2}), one transmission gate ($M_{U3,4}$), and logics. The A-FIFO$_U$ consists of 8 C_U (22.8 fF$\times$8), three switches (M_{Q0-2}), logics, and a differential pair (M_{Q3-6}) for independent analog-to-digital conversion in each unit for the compression mode. The node voltage, $V_{A\text{-FIFO}}$, is connected to the same nodes in other A-FIFO$_U$ and plays the common node for the normal mode operation by MOD (mode change) signal. The differential pair converts the analog value of the connected A-FIFO$_U$ by comparing V_{REF} or V_{RAMP} according to the mode of operation and delivers the comparison result (I_{OUT}, I_{OUTB}) to the cross-coupled pair in the dynamic comparator as shown in Fig. 13. The cross-coupled pair is a common circuit for the 16 differential pairs. For the matching of capacitors, we place dummies on top and bottom sides (y-direction) and one for rectangular shape, as shown in Fig. 12. The matching in the x-direction is not a concern because each channel was placed as column parallel in the layout.

E. Integrated Circuits in Digital Domain

The LFPs are temporally compressed via Δ_T-modulation, and then digitized by $\Delta\Sigma$-ADC and sent to an on-chip decimation filter. After the decimation, the LFPs are subtracted (spatial difference, Δ_S) channel-by-channel according to the external programming, which is determined by the information based on the physical configuration of the electrodes in probes to maximize the spatial correlation. Since our 128-channel recording interface chip will be tested for a neural probe having eight shanks and 16 recording sites in each shank, Δ_S is taken within a subgroup (16-channel). In this implementation, Δ_S is achieved in two ways: 1) taking differences among adjacent channels or 2) taking center-around differences, as shown in Fig. 14. Once the one out of 16 channels is

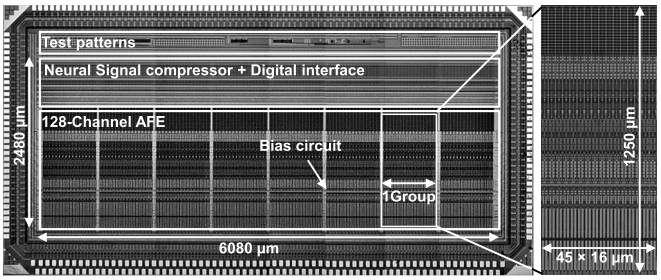


Fig. 15. Die microphotograph of the fabricated neural recording/compression interface. A group (16-channel) is enlarged on the right side.

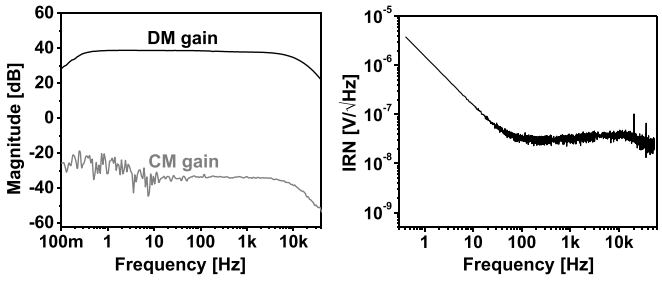


Fig. 16. (a) Measured frequency response of the LNA. (b) Measured IRN spectral density of the LNA.

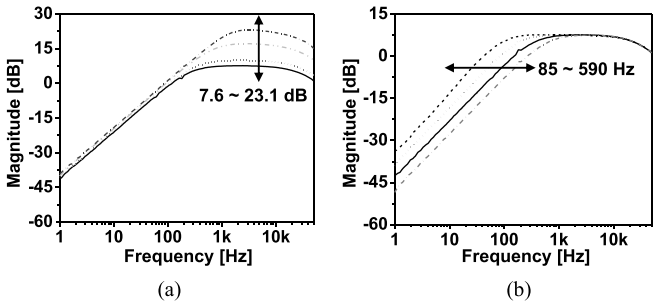


Fig. 17. Measured frequency response of the HPF. (a) Different gain sets and (b) different frequency sets.

designated as a reference (x_{ref}), we have two options: 1) taking differences between the adjacent channels and x_{ref} or 2) taking differences between x_{ref} and the rest of the channels. The cases, 1) and 2), are mathematically given by

$$\Delta_n = x_{\text{mod }(n+n_{ref},N)} - x_{\text{mod }(n+1+n_{\text{ref}},N)} \tag{7a}$$

$$\Delta_n = x_{n_{\text{ref}}} - x_{\text{mod }(n+n_{\text{ref}},N)} \tag{7b}$$

where $N = 16$, $n = 0, 1 \ldots 15$, integers. To effectively use the spatial correlation, (7a) [Fig. 14(a)] may be the best choice since Δ_S is made with the adjacent channels. However, it is prone to failure because the reconstruction is impossible if only one electrode out of 16 is not properly working. Thus, Fig. 14(b) was prepared as an alternative because we can retrieve signals, even in the case when some of the channels out of 16 are not operating. The selection for the reference and Δ_S with (a) or (b) is externally programmable.

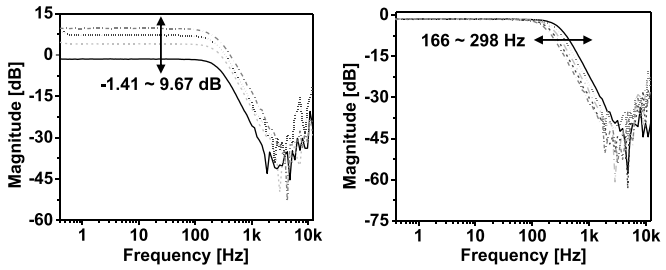
The Huffman encoders, decimation filters, data serializers, and spike frame counters are also implemented in the digital domain. In particular, a 2nd-order *sinc* filter was realized to decimate the output bit stream from the $\Delta_T - \Delta\Sigma$ ADC [41]. Based on the prior investigation of the pre-recorded LFPs, the codebook for the Huffman encoder was prepared to cover 5-b resolution, but it is also externally programmable to reduce the size of the headers in the encoder when applicable. For spikes, 8-b programmable frame counters are implemented in each channel to adjust the length of the extracted spike waveforms.

V. Experimental Results

The proposed modular interface prototype chip was fabricated in 180-nm 1P6M CMOS processes. Fig. 15 shows a microphotograph of the fabricated chip. The overall die area is 6080×2480 mm^2 excluding test patterns and pads. As indicated, the 128-channel recording circuits are placed in column parallel, and all the controllers for SAR and SS ADCs, frame counters for spikes, and Huffman encoders and decimation filters, Δ_S generation circuit for LFPs are placed on top of the 128 channels. The channel pitch is 45 μm with a physical separation of 5 μm to suppress the channel crosstalk. The measured channel crosstalk is negligible (< -100 dB), since the physical channel separation is relatively large, compared to the channel pitch and each channel independently operates without analog multiplexing. The digital circuits occupy 6080×930 μm^2. The 128-channel recording circuit takes an area of 6080×1250 μm^2 including eight bias circuits for 8×16 channels.

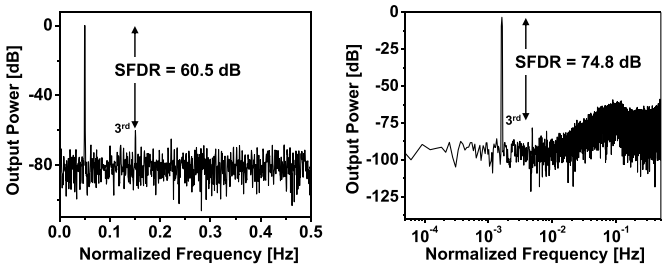
A. Performance Measurement

The measured differential and common mode gains of the LNA are shown in Fig. 16(a). The LNA has a mid-band gain of 37.8 dB from 0.4-Hz low to 9-kHz high-frequency corner, and ~60-dB CMRR at the mid-band. The measured IRN of the LNA is also shown in Fig. 16(b). The integrated IRN from 0.4 Hz to 10.3 kHz is ~5.18 μV_{rms}, satisfying the noise requirement of the neural recording AFE (<10 μV_{rms}) [26]. The total harmonic distortion (THD) of the LNA with 1 kHz sine wave input was also measured. The THD was about -64.1 dB that is equivalent to ~0.062% with 3.2 mV$_{\text{pp}}$ input. The power consumption of the LNA was 1.045 μW. Based on the foregoing measurements, the performance metrics for the LNA such as NEF and NEF$^2 V_{\text{DD}}$ are calculated as 2.56 and 3.28, respectively. The measured frequency responses of the HPF for spikes are shown in Fig. 17. There are four levels of the voltage gains of 7.6, 9.9, 17.1, and 23.1 dB [Fig. 17(a)]. The f_{HPF} adjustment of the HPF is shown in Fig. 17(b) as well. It varies from 85 to 590 Hz by setting the current DAC externally. The frequency responses of the SC-LPF were also measured as shown in Fig. 18. The LPF has four different gain settings: -1.41, 2.35, 7.42, and 9.67 dB [Fig. 18(a)] and four different f_L adjustment from 166 to 298 Hz [Fig. 18(b)] There is no bump in the response thanks to the low Q_{LPF}. Fig. 19(a) shows an fast Fourier transform (FFT) analysis of the SAR ADC with a 1-KHz sine wave. The spurious free dynamic range (SFDR) is ~60.5 dB and



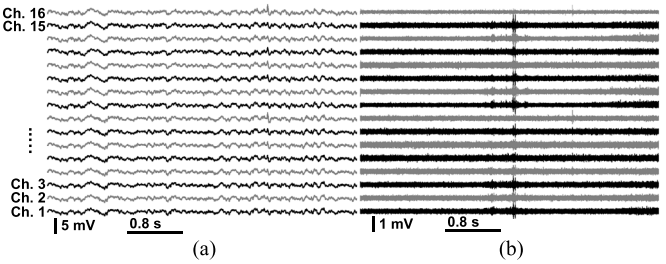
(a) (b)

Fig. 18. Measured frequency response of the LPF. (a) Different gain sets and (b) different frequency sets.



(a) (b)

Fig. 19. FFTs of the output of (a) SAR and (b) $\Delta - \Delta\Sigma$ ADCs.



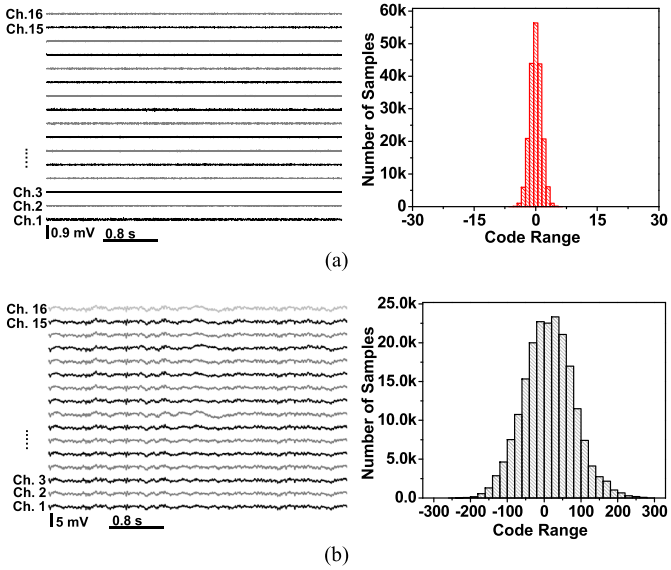
(a) (b)

Fig. 20. 5-s *in vivo* data from multi-channel recording. (a) LFPs. (b) Spikes

the signal to noise and distortion ratio (SNDR) is 48.76 dB, equivalent to the 7.81 b effective number of bit (ENOB). The power consumption of the SAR ADC is ~0.11 μW and its figure of merit (FoM) is 19.61 fJ/C-s. The max/min differential nonlinearity (DNL) and integral nonlinearity (INL) were also measured. Those were 0.629/−0.4182 for DNL and 0.465/−0.511 for INL. Fig. 19(b) shows the FFT result of the $\Delta - \Delta\Sigma$ ADC with a 100 Hz sine wave after the reconstruction process. From Fig. 19(b) the SFDR is ~74.8 dB, and the calculated SNDR is 61.87 dB, equivalent to 9.98-b ENOB. The power consumption of the $\Delta - \Delta\Sigma$ ADC is measured as ~0.837 μW at 64 kS/s and its FoM is 414.15 fJ/C-s.

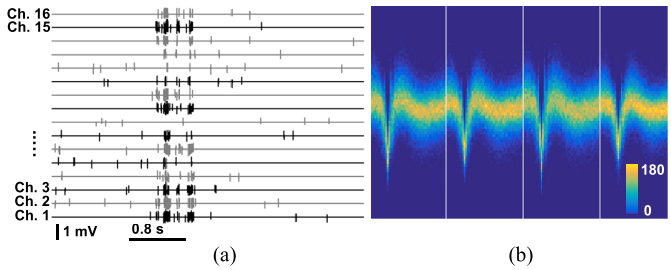
B. In Vivo Measurement

For *in vivo* validation, a 32-channel tetrode was implanted in the neo-cortex region of the brain in a rat and externally interfaced with the fabricated chip by using short (<20 cm), flexible wires. The animal under measurement was not anesthetized to observe the activity-dependent power consumption of the digital circuits in real time. The interface circuit was mainly powered through wires and different supplies necessary for this circuit were generated on-board. The data are reconstructed at the off-chip host computer by using a commercial software.



(a) (b)

Fig. 21. 5-s *in vivo* multi-channel LFP data. (a) Compressed (b) reconstructed ones and their histogram.



(a) (b)

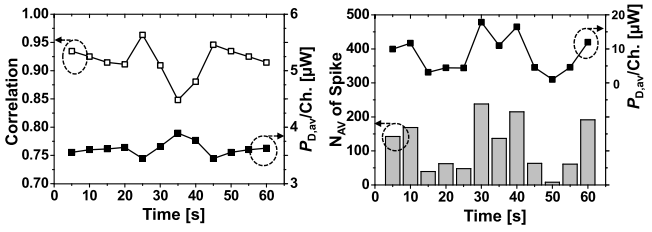
Fig. 22. 5-s *in vivo* spikes in the compression mode. (a) Transient waveforms. (b) Aligned spikes represented in 2-D density.

Fig. 20(a) and (b) show ~5-s snapshots of 16 channels: reconstructed LFPs and spikes in the normal mode, respectively, with the lowest gains and $f_{\mathrm{LPF}} = 248$ Hz and $f_{\mathrm{HPF}} = 450$ Hz. Because of the normal mode setting, the spikes and ISIs are plotted together in Fig. 20(b). Fig. 21(a) and (b) depicts ~5-s LFPs (different *in vivo* sessions from Fig. 20) before and after reconstruction, respectively. Because it is hard to judge the effectiveness of the compression by only observing the transient waveforms of the LFPs, the histograms of the code distributions for both (a) and (b) are also presented, while the codes in Fig. 21(b) spans from −249 to 225, those in Fig. 21(a) was distributed only from −7 to 6. Fig. 22(a) shows the extracted waveforms of spikes in the compression mode. For the spikes, only the active parts have the codes, otherwise, all data remain constant ("0" in those cases). The alignment of the extracted spikes has been done using an open-source software (UltraMegaSort 2000) and the aligned spikes from four different channels in 2-D density are shown in Fig. 22(b) [42]. We used 4 ms for the spike frame counter setting for this measurement. Fig. 23 shows the power consumption (P_D) in the digital circuits when the compression was performed. Fig. 23(a) shows the 1 min average power consumption per channel ($P_{D,\mathrm{av}}$/Ch) measurement and the spatial-correlation based on the measured LFP. As shown, the overall $P_{D,\mathrm{av}}$/Ch follows well with the correlation between channels: it means the compression based on the correlation is

TABLE I
PERFORMANCE SUMMARY AND COMPARISON

	[3]	[4]	[28]	[29]	**This work**
Analog power/Ch. [μW]	49.06	27.84	7.7	42.5	**3.37**
Digital power/Ch. [μW]			43	95.6	**11.98**
Analog area/Ch. [mm^2]	0.12	0.19	–	[1)]0.144	**0.059**
Digital area/Ch. [mm^2]			–	[1)]0.116	**0.039**
Data comp. for LFP/spike	No/No	No/No	No/Yes	–/Yes	**Yes/Yes**
Lossless LFP/spike	–/–	–/–	–/Yes	–/Yes	**Yes**
f_S for LFP/spike [KHz]	2.5/30	2.5/30	1/20	–/30	**2/25**
Bits/Sample for LFP/spike	10	10	–	–	**11/8**
ENOB for LFP/spike [bit]	–	9.2	> 9	–/7	**9.98/7.81**
FoM for LFP/spike [fJ/C-s]	–	[2)]397.9	–/–	–/[2)]1927.1	**414.15/19.61**
IRN [μV_{rms}]	6.36	3.2	3.3/3.1	5.9	**5.18**
Bandwidth [kHz]	10	[1)]6	10	9	**9.2**
NEF	–	3.08	3.8	4.9	**2.61**
NEF^2VDD	–	17.13	–	[1)]6.73	**3.41**
CMRR [dB]	> 60	60	> 70	–	**60**
# of chan. for LFP/spike	384(966)	52 (455)	6/24	–/16	**128/128**
Supply voltage [V]	1.8	1.8	–	1.2	**0.5/1.0 for analog 1.0/1.8 for digital**
Technology	130 nm SOI	180 nm	180 nm	180 nm	**180 nm**

[1)] Estimated. [2)] Calculated based on 7.4 μW of power consumption in ADC.



(a) (b)

Fig. 23. (a) Power consumption in digital circuits for LFPs and spatial correlation of the LFPs. (b) Power consumption in digital circuits for spikes and number of detected spikes.

effective. The $P_{D,\text{av}}$/Ch for LFPs is measured as ∼3.59 μW. Fig. 23(b) shows another $P_{D,\text{av}}$/Ch measurement for spike recording. As anticipated, it is proportional to the spike firing rate; it is activity dependent. The $P_{D,\text{av}}$/Ch for spikes is ∼8.39 μW.

C. Performance Comparison

The overall performance of the fabricated neural recording interface with lossless compression is summarized in Table I by comparing with recent state-of-the-art works. The measured power consumption of the AFE and related digital circuits in the compression mode are about 3.37 and 11.98 μW/Ch, respectively. This paper achieved the smallest power consumption per channel while recording simultaneously LFPs and spikes. At the same time, the proposed work accomplished high quality recording in terms of IRN and resolutions (10 b for LFPs, 8 b for spikes). Considering the P_D is 107.5 μW/Ch in the uncompressed case, the dynamic power reduction in the digital circuit is significant, estimated as ∼1/9.

VI. CONCLUSION

We report a scalable128-channel neural recording interface chip with embedded lossless compression to reduce the dynamic power consumption for data transmission in high-density neural recording system. We utilized the inherent characteristics of neural signals: spatiotemporal correlation of LFPs and temporal sparsity of spikes. We have demonstrated that effective on-chip lossless neural signal compression can be achieved while performing high-fidelity recording. From the implemented compression scheme, the data rate of neural signals could be reduced by a factor of 5.35 for LFPs and 10.54 for spikes, respectively, resulting in dynamic power reduction by 89% (107.5 to 11.98 μW/Ch) compared to the uncompressed case. The fabricated circuits also demonstrated the state-of-the-art performance of 3.37 μW/Ch, 5.18 μV_{rms} noise, and 3.41 $\text{NEF}^2 V_{\text{DD}}$.

ACKNOWLEDGMENT

The authors would like to thank Dr. A. Mohebi and Prof. J. Berke, Department of Neurology, University of California, San Francisco, CA, USA, for providing the *in vivo* measurement environment. We would also like to thank Prof. G. Buzsáki, School of Medicine, New York University, New York, NY, USA, for allowing us to use the pre-recorded neural signal from his laboratory in our analysis.

REFERENCES

[1] I. H. Stevenson and K. P. Kording, "How advances in neural recording affect data analysis," *Nature Neurosci.*, vol. 14, no. 2, pp. 42–139, Feb. 2011.

[2] B. C. Raducanu *et al.*, "Time multiplexed active neural probe with 678 parallel recording sites," in *Proc. Eur. Solid-State Device Res. Conf.*, Sep. 2016, pp. 385–388.

[3] C. M. Lopez *et al.*, "A 966-electrode neural probe with 384 configurable channels in 0.13 μm SOI CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Jan. 2016, pp. 392–393.

[4] C. M. Lopez *et al.*, "An implantable 455-active-electrode 52-channel CMOS neural probe," *IEEE J. Solid-State Circuits*, vol. 49, no. 1, pp. 248–261, Jan. 2014.

[5] G. Buzsáki, "Large-scale recording of neuronal ensembles," *Nature Neurosci.*, vol. 7, no. 5, pp. 446–451, May 2004.

[6] G. Buzsáki, C. A. Anastassiou, and C. Koch, "The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes," *Nature Rev. Neurosci.*, vol. 13, no. 6, pp. 407–420, Jun. 2012.

[7] H. Chandrakumar and D. Marković, "A high dynamic-range neural recording chopper amplifier for simultaneous neural recording and stimulation," *IEEE J. Solid-State Circuits*, vol. 52, no. 3, pp. 645–656, May 2017.

[8] W. Jiang, V. Hokhikyan, H. Chandrakumar, V. Karkare, and D. Marković, "A ±50-mV linear-input-range VCO-based neural-recording front-end with digital nonlinearity correction," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 173–184, Jan. 2017.

[9] W. Biederman *et al.*, "A 4.78 mm^2 fully-integrated neuromodulation SoC combining 64 acquisition channels with digital compression and simultaneous dual stimulation," *IEEE J. Solid-State Circuits*, vol. 50, no. 4, pp. 1038–1047, Apr. 2015.

[10] K. A. Ng and Y. P. Xu, "A low-power, high CMRR neural amplifier system employing CMOS inverter-based OTAs with CMFB through supply rails," *IEEE J. Solid-State Circuits*, vol. 51, no. 3, pp. 724–737, Mar. 2015.

[11] S.-Y. Park, J. Cho, K. Na, and E. Yoon, "Toward 1024-channel parallel neural recording: Modular Δ-$\Delta\Sigma$ analog front-end architecture with 4.84 fJ/C-s·mm^2 energy-area product," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2015, pp. 112–113.

[12] D. Han, Y. Zheng, R. Rajkumar, G. S. Dawe, and M. Je, "A 0.45 V 100-channel neural-recording IC with sub-μW/channel consumption in 0.18 μm CMOS," *IEEE Trans. Biomed. Circuits Syst.*, vol. 7, no. 6, pp. 735–746, Dec. 2013.

[13] H. Kassiri *et al.*, "All-wireless 64-channel 0.013 mm^2/ch closed-loop neurostimulator with rail-to-rail DC offset removal," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 452–454.

[14] R. Muller, S. Gambini, and J. M. Rabaey, "A 0.013 mm^2, 5 μW, DC-coupled neural signal acquisition IC with 0.5 V supply," *IEEE J. Solid-State Circuits*, vol. 47, no. 1, pp. 232–243, Jan. 2012.

[15] S.-I. Chang, S.-Y. Park, and E. Yoon, "Low-power low-noise pseudo-open-loop preamplifier for neural interfaces," *IEEE Sensors J.*, vol. 17, no. 15, pp. 4843–4852, Aug. 2017.

[16] S.-Y. Park, J. Cho, and E. Yoon, "3.37 μW/Ch modular scalable neural recording system with embedded lossless compression for dynamic power reduction," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2017, pp. 168–169.

[17] J. Xu, T. Wu, W. Liu, and Z. Yang, "A frequency shaping neural recorder with 3 pF input capacitance and 11 plus 4.5 bits dynamic range," *IEEE Trans. Biomed. Circuits Syst.*, vol. 8, no. 4, pp. 510–527, Aug. 2014.

[18] W. A. Smith, B. J. Mogen, E. E. Fetz, V. S. Sathe, and B. P. Otis, "Exploiting electrocorticographic spectral characteristics for optimized signal chain design: A 1.08 W analog front end with reduced ADC resolution requirements," *IEEE Trans. Biomed. Circuits Syst.*, vol. 10, no. 6, pp. 1171–1180, Dec. 2016.

[19] J. Scholvin *et al.*, "Close-packed silicon microelectrodes for scalable spatially oversampled neural recording," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 1, pp. 120–130, Jan. 2016.

[20] S. M. S. Jalaleddine, C. G. Hutchens, R. D. Strattan, and W. A. Coberly, "ECG data compression techniques—A unified approach," *IEEE Trans. Biomed. Eng.*, vol. 37, no. 4, pp. 329–343, Apr. 1990.

[21] G. Antoniol and P. Tonella, "EEG data compression techniques," *IEEE Trans. Biomed. Eng.*, vol. 44, no. 2, pp. 105–114, Feb. 1997.

[22] D. Staudenmann, I. Kingma, A. Daffertshofer, D. F. Stegeman, and J. H. van Dieen, "Improving EMG-based muscle force estimation by using a high-density EMG grid and principal component analysis," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 4, pp. 712–719, Apr. 2006.

[23] L. Brechet, M.-F. Lucas, C. Doncarli, and D. Farina, "Compression of biomedical signals with mother wavelet optimization and best-basis wavelet packet selection," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 12, pp. 2186–2192, Dec. 2007.

[24] E. S. G. Carotti, J. C. de Martin, R. Merletti, and D. Farina, "Compression of multidimensional biomedical signals with spatial and temporal codebook-excited linear prediction," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 11, pp. 2604–2610, Nov. 2009.

[25] Y. Wongsawatt, S. Oraintarat, T. Tanakat, and K. R. Rao, "Lossless multi-channel EEG compression," in *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 6. May 2006, p. 1614.

[26] T. Jochum, T. Denison, and P. Wolf, "Integrated circuit amplifiers for multi-electrode intracortical recording," *J. Neural Eng.*, vol. 6, no. 1, p. 12001, Jan. 2009.

[27] R. Q. Quiroga, Z. Nadasdy, and Y. Ben-Shaul, "Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering," *Neural Comput.*, vol. 16, no. 8, pp. 1661–1687, Aug. 2004.

[28] S. Mitra, J. Putzeys, C. M. Lopez, C. M. A. Pennartz, and R. F. Yazicioglu, "24 channel dual-band wireless neural recorder with activity-dependent power consumption," *Analog Integr. Circuits Signal Process.*, vol. 83, no. 3, pp. 317–329, 2015.

[29] B. Gosselin *et al.*, "A mixed-signal multichip neural recording interface with bandwidth reduction," *IEEE Trans. Biomed. Circuits Syst.*, vol. 3, no. 3, pp. 129–141, Jun. 2009.

[30] B. Gosselin and M. Sawan, "Adaptive detection of action potentials using ultra low-power CMOS circuits," in *Proc. IEEE Biomed. Circuits Syst. Conf.*, Nov. 2008, pp. 209–212.

[31] R. R. Harrison, "A low-power integrated circuit for adaptive detection of action potentials in noisy signals," in *Proc. IEEE Eng. Med. Biol. Soc.*, vol. 4. Sep. 2003, pp. 3325–3328.

[32] B. Gosselin and M. Sawan, "An ultra low-power CMOS automatic action potential detector," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 17, no. 4, pp. 346–353, Aug. 2009.

[33] R. R. Harrison and C. Charles, "A low-power low-noise CMOS amplifier for neural recording applications," *IEEE J. Solid-State Circuits*, vol. 38, no. 6, pp. 958–965, Jun. 2003.

[34] S.-Y. Park, J. Cho, K. Na, and E. Yoon, "Modular 128-channel Δ-$\Delta\Sigma$ analog front-end architecture using spectrum equalization scheme for 1024-channel 3-D neural recording microsystems," *IEEE J. Solid-State Circuits*, to be published, doi: 10.1109/JSSC.2017.2764053.

[35] F. Zhang, J. Holleman, and B. P. Otis, "Design of ultra-low power biopotential amplifiers for biosignal acquisition applications," *IEEE Trans. Biomed. Circuits Syst.*, vol. 6, no. 4, pp. 344–355, Aug. 2012.

[36] C. C. Enz, F. Krummenacher, and E. A. Vittoz, "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Analog Integr. Circuits Signal Process.*, vol. 8, no. 1, pp. 83–114, Jul. 1995.

[37] S. Chang, "Nano-watt modular integrated circuits for wireless neural interface," Univ. Michigan, Ann Arbor, MI, USA, 2013.

[38] J. Roh, S. Byun, Y. Choi, H. Roh, Y.-G. Kim, and J.-K. Kwon, "A 0.9-V 60-μW 1-bit fourth-order delta-sigma modulator with 83-dB dynamic range," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 361–370, Feb. 2008.

[39] X. Zou, X. Xu, L. Yao, and Y. Lian, "A 1-V 450-nW fully integrated programmable biomedical sensor interface chip," *IEEE J. Solid-State Circuits*, vol. 44, no. 4, pp. 1067–1077, Apr. 2009.

[40] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, Oct. 1989.

[41] J. Candy, "Decimation for sigma delta modulation," *IEEE Trans. Commun.*, vol. 34, no. 1, pp. 72–76, Jan. 1986.

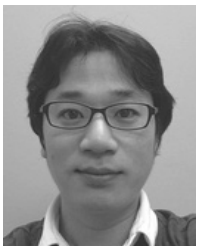
[42] D. N. Hill, S. B. Mehta, and D. Kleinfeld. (2012). *Ultra-MegaSort 2000 Manual*. [Online]. Available: https://neurophysics.ucsd.edu/lab/UltraMegaSort2000%20Manual.pdf

Sung-Yun Park received the B.S. degree in electrical engineering from Pusan National University, Busan, South Korea, in 2005, and the M.S. degree in electrical engineering from the Korean Advanced Institute of Science and Technology, Daejeon, South Korea, in 2008, where he was involved in photonic device research. He continued his study in integrated circuits, Cornell University, Ithaca, NY, USA, and the M.Eng. degree in electrical and computer engineering and the Ph.D. degree in electrical engineering with the University of Michigan, Ann Arbor, MI, USA, in 2011 and 2016, respectively.

From 2008 to 2010, he was with the Fairchild Semiconductor, Bucheon, South Korea, as a Member of Technical Staff for high-voltage mixed-signal integrated circuit design. He is currently a Research Fellow with the University of Michigan. His current research interests include low power, low noise mixed-signal integrated circuits and power management circuits for biomedical systems.

Jihyun Cho received the B.S. and M.S. degrees in electrical engineering from Yonsei University, Seoul, South Korea, in 2005 and 2007, respectively. From 2007 to 2010, he was with the Republic of Korea Air Force Academy, Cheongwon, South Korea, as a Lecturer. He received the Ph.D. degree in electrical engineering from the University of Michigan, Ann Arbor, MI, USA.

He is currently with Apple Inc., Cupertino, CA, USA, as a System Designer. His current research interests include CMOS image sensors and mixed-signal VLSI circuit designs.



Kyuseok Lee received the B.S. degree in electrical engineering from Yonsei University, Seoul, South Korea, in 2005. He is currently pursuing the Ph.D. degree with the University of Michigan, Ann Arbor, MI, USA.

From 2005 to 2012, he was with SK Hynix, Icheon, South Korea, where he is involved in high-speed DRAM for GPU, test structures for yield analysis, and CMOS image sensors. His current research interests include smart, low-power CMOS image sensors and neuromorphic mixed-signal processing units for a vision-based navigation system in autonomous aviation applications.

Euisik Yoon (M'82) received the B.S. and M.S. degrees in electronics engineering from Seoul National University, Seoul, South Korea, in 1982 and 1984, respectively, and the Ph.D. degree in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 1990.

From 1990 to 1994, he was with the National Semiconductor Corp., Santa Clara, CA, USA. From 1994 to 1996, he was the Technical Staff Member at Silicon Graphics Inc., Mountain View, CA, USA. He took faculty positions in the Department of Electrical Engineering at the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, during 1996–2005 and in the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA, during 2005–2008, respectively. During the academic year of 2000–2001, he was a Visiting Faculty at the Agilent Laboratory, Palo Alto, CA, USA. In 2008, he joined the Department of Electrical Engineering and Computer Science at the University of Michigan where he is a Professor and the Director of the NSF International Program for Advancement of Neurotechnology (IPAN). His current research interests include MEMS, integrated microsystems, and VLSI circuit design.

Dr. Yoon has served on various Technical Program Committees including the Microprocesses and Nanotechnology Conference 1998, the International Sensor Conference 2001, the IEEE Asia–Pacific Conference on Advanced System Integrated Circuits 2001 and 2002, the International Conference on Solid-State Sensors, Actuators and Microsystems (Transducers) 2003, 2005, the IEEE International Electron Device Meeting 2006–2008 and the IEEE International Conference on Micro Electro Mechanical Systems 2006, 2009, and 2010. He also served on the IEEE International Solid-State Circuit Conference program committee 2003–2007 and was a General Chair of International Symposium on Bio Micro & Nanosystems 2005. Currently he serves as an an associate editor for IEEE SOLID-STATE CIRCUITS LETTERS.