Neural Spike Compression through Salient Sample Extraction and Curve Fitting Dedicated to High-Density Brain Implants

Mahdi Nekoui Shahraki

A DISSERTAION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

YORK UNIVERSITY

TORONTO, ONTARIO

September 2023

Abstract

This work proposes a data reduction framework, specific to the compression of extra-cellular neuronal action potentials on brain-implantable microsystems. The proposed framework significantly reduces the extent of data representing spike waveforms, paving the way for the implementation of next-generation, high-density neural recording brain implants. This highly-compressive approach picks a small number of salient samples of the spike, using which and based on some predefined functions the entire spike waveshape is formulated. The amplitudes and timings of the salient samples are sent off the implant in order to reconstruct the spike waveshape on the external side of the system. In addition to exhibiting extremely high data compression capability, this technique is highly hardware efficient, hence it well suits for brain-implantable neural recording microsystems with high channel counts. Based on the proposed framework, a 128-channel neural signal compressor is designed and microfabricated using the TSMC 130-nm CMOS technology, and measures 1.05 mm by 0.35 mm, giving an area-per-channel of 0.00287 mm^2. The circuit is tested using a library of intra-cortically recorded neural signals. At an average spike firing rate of 8 Spike/s, the circuit temporally reduces neural data with an average compression rate of ~272, which is equivalent to a true compression rate of ~2176. Operated using a 1-V power supply and at a clock rate of 32 MHz, the 128-channel neural data compressor consumes 0.164 μW/channel.

Dedication

This thesis is dedicated to my deceased mother.

"Dear mom, you will forever remain in my heart, even as tears flow for the life you lived. It's still difficult for me to accept that you're gone..."

Acknowledgments

I would like to express my truthful gratitude to my supervisor Professor Amir M. Sodagar for his dedication to help me. His advice helped me to pass the obstacles I had confronted and accomplish this part of my life.

I also would like to thank my beloved wife, Zahra, children, Noora and Iliya, for their continuous encouragement through my whole life. To those who indirectly contributed in this research, your kindness means a lot to me. Thank you very much.

I would like to thank Professors Hossein Kassiri and Sebastian Magierowski for helping to enhance my thesis with their valuable comments.

Finally, I appreciate my friends', Mohammad Ali Shaeri, Mansour Taghadosi, Alireza Dabaghian, Tayebeh Yousefi, Milad Akbari, and Hamid Taheri, encouragement and support during these five years for completing this task.

Contents

List of Tables

List of Figures

Chapter 1 Introduction to Signal Processing on Brain Implants

1.1 Introduction

Brain-implantable microsystems are sophisticated devices specifically engineered to interface with the intricate neural networks present within the brain. By residing directly inside or in close proximity to the brain, these systems open up possibilities for highly precise and detailed neural interfacing. The proximity to the neural tissue enables capturing neural activities with high temporal and spatial resolutions, providing highly-informative insight into the functioning of the brain [1].

As depicted in Figure 1-1, the implanted device plays a crucial role in facilitating communication with an external host. This external host serves two main purposes: configuration of the implantable microsystem and exchange of signals. Configuring the device involves setting parameters, adjusting settings, and calibrating it to suit the specific requirements of the individual

user. Signal exchange refers to the transfer of neural data recorded by the implant to the external system, enabling further analysis, interpretation, and utilization of the recorded information [2].

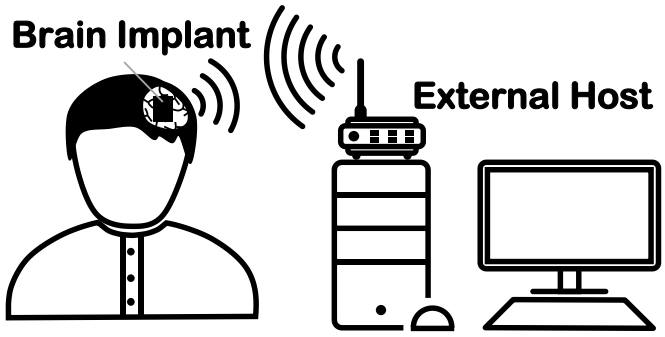


Figure 1-1: Implantable brain-implantable microsystems for high-resolution neural interfacing [3].

One of the primary functions of brain-implantable devices is the recording and streaming of neuronal activities to the outside world. Through a wireless connection, these devices enable real-time monitoring and analysis of neural signals, providing valuable data for research, medical diagnostics, and therapeutic applications. The ability to wirelessly transmit neural data eliminates the need for cumbersome physical connections and allows for greater freedom of movement and comfort for the user.

The recording and streaming of neuronal activities have numerous applications and benefits. In the field of neuroscience, it provides researchers with invaluable insights into brain functions, neural pathways, and the underlying mechanisms of various neurological disorders. For clinical

purposes, brain-implantable microsystems offer the potential for personalized medicine, as they enable continuous monitoring and analysis of brain activity in patients with conditions such as epilepsy, Parkinson's disease, or paralysis. This real-time data can inform treatment strategies and facilitate the development of targeted therapies.

Furthermore, brain-implantable microsystems contribute to advancements in the field of cognitive sciences. By capturing and analyzing neural signals associated with cognitive processes, these devices aid in understanding human cognition, memory formation, and decision-making processes. They also have the potential to enhance brain-machine interfaces (BMI), enabling individuals to control external devices or prosthetics using their neural activity.

In summary, brain-implantable microsystems offer a remarkable opportunity to interface with the brain at a high resolution. They facilitate the recording and transmission of neural activities, enabling real-time monitoring, analysis, and applications in diverse fields such as neuroscience, medicine, and cognitive sciences. Through wireless connectivity, these devices provide a means for exchanging data between the implanted system and external hosts, unlocking new possibilities for research, diagnostics, and therapeutic interventions [1], [2].

1.2 General Overview

In the last two decades, advancements in neural interfacing devices have transformed from microfabricated microelectrode arrays to fully-implantable microsystems that function independently and establish wireless connections with the outside world [4], [5]. These brain-implantable microsystems have expanded their applications from fundamental and clinical neuroscience to include neuro-prostheses, therapeutic treatments, cognitive science research and development, and brain-machine interfacing (BMI). Figure 1-2 illustrates a simplified functional diagram of the implantable module within an intra-cortical neural recording system, commonly known as a brain implant. This module utilizes a microelectrode array to capture neural signals, which are then pre-amplified and filtered in the analog domain. Subsequently, the recorded signals are digitized to leverage digital signal processing techniques when necessary and employ digital communication methods for transmitting the recorded data to an external module. Collectively, the microelectrode array, analog signal preconditioning circuits, and analog-to-digital converter are commonly referred to as the recording front-end of the system.

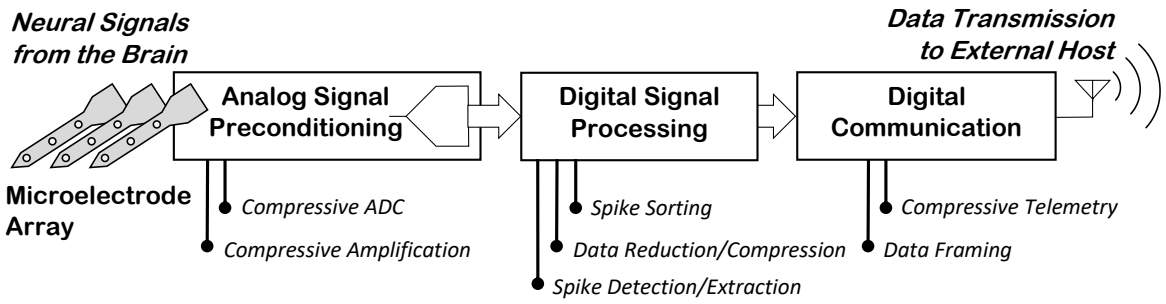


Figure 1-2: Signal processing opportunities along the signal travel path in a neural recording brain implant [6].

1.2.1 Signals of Interest

In the neuronal network of the brain, information is encoded by modifying the pattern of spiking activities known as "action potentials" or simply "spikes." These all-or-none spiking activities, resulting from rapid changes in cell membrane voltage, play a crucial role in transmitting information through electrochemical signaling in the central nervous system [7].

To closely monitor intra-cortical spiking activities near neurons, an electrode array can be used, providing high temporal and spatial resolution [8], [9], [10], [11], [12], [13]. These activities find applications in prosthetics and rehabilitation.Figure 1-3 illustrates the major components of an extra-cortically recorded neural signal: "action potentials," "local field potentials (LFPs)," and "background noise." Action potentials refer to electrochemical impulses generated by neurons in response to input from other neurons. Recorded electrically near the firing neuron, an extracellular

action potential exhibits amplitude fluctuation ranging from tens to hundreds of microvolts, typically lasting 1 to 2 milliseconds. LFPs, on the other hand, represent low-frequency signal components believed to be an average of neuronal activities occurring at a distance. While some researchers extract relevant information from LFPs, action potentials remain the primary source of information in intra-cortical neural signals. Additionally, neural signals are often contaminated by background noise, which degrades signal quality and hampers signal processing performance. It should be noted, firstly, the utilization of spikes extends beyond clinical applications. Neuroscientists, for instance, are deeply engaged in the study of spikes. Secondly, the attributes of a spike are defined based on the specific application. In prosthetic applications, for instance, the focus might solely be on the occurrence of the spike. Meanwhile, certain neuroscientists are intrigued by the origin of the spike. In such cases, the firing neuron is identified through spike sorting methods. To maintain the high spatial resolution of action potentials during intra-cortical recording and simultaneously eliminate LFPs and other undesirable low-frequency components (such as DC offset and signal preconditioning circuit drift), neural signals undergo preamplification and filtering. This process involves applying a low cut-off frequency of 100-300 Hz and a bandwidth of 6000-10000 Hz. Subsequently, to facilitate digital handling and processing, the signals are sampled at a rate of up to 20-30 kSamples/second and quantized with a typical resolution of 8-10 bits [1].

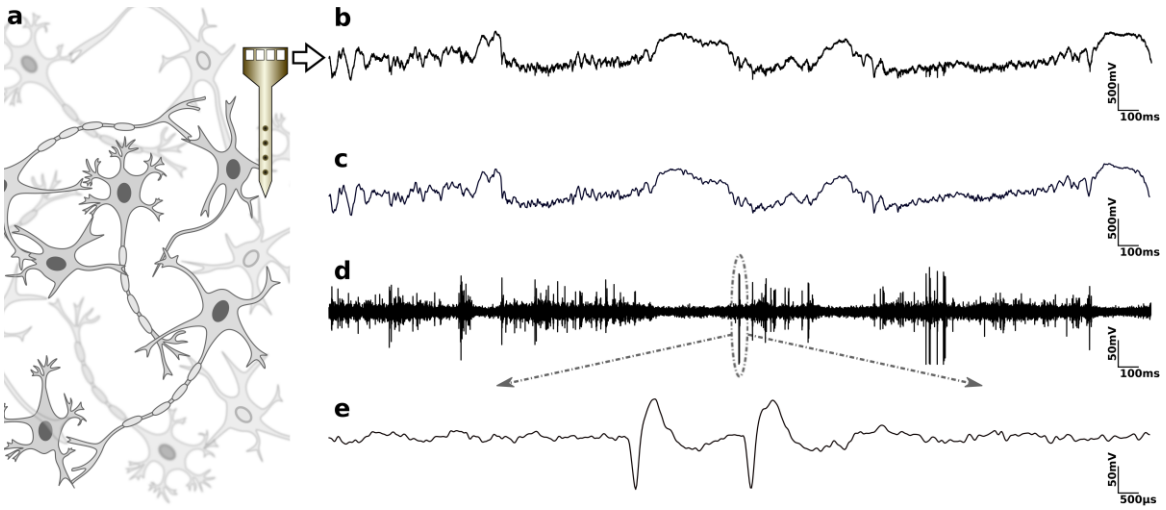


Figure 1-3: An extra-cellular neuronal signal: (a) recording of extra-cellular activities from a neuronal population; (b) the raw signal; (c) the LFP component (<150 Hz); (d) mixture of spikes (300 to 6000 Hz) and background noise; and (e) close-up view of the signal shown in (d). The neuronal signal shown in this figure is recorded from the IT cortex of adult male rhesus monkeys [14]. (Figure from: [2])

1.2.2 Functions and Applications

The complexity of the human brain necessitates a comprehensive recording of neural signals across multiple regions. Neuroscientists require high-density recordings to ensure they capture the distributed information within the brain network. These recorded signals are then subjected to careful processing to extract meaningful content while filtering out task-irrelevant information

[15]. This processing step helps researchers uncover valuable insights into brain functioning and cognitive processes.

One important aspect of advancing neural interfaces (NIs) is on-chip processing, which contributes to the efficiency and effectiveness of the system. By performing data processing directly on the neural interface chip, the volume of recorded data can be significantly reduced. This reduction not only minimizes the storage requirements but also enables efficient data transfer with lower transmission power. The integration of on-chip processing capabilities enhances the overall performance of NIs, making them more practical and viable for real-time brain signal analysis.

Looking towards the future, NIs are being designed with specific applications in mind, particularly in the realms of prosthetics and therapeutics. Prosthetic NIs are envisioned as portable and hardware-efficient devices that aim to restore lost brain functions, such as movement and speech. These innovative interfaces hold the potential to assist individuals with disabilities by providing them with the means to control prosthetic limbs or communicate through neural signals.

Furthermore, advanced NIs capable of both neural recording and stimulation are paving the way for closed-loop systems. These closed-loop NIs detect neuromarkers associated with various brain disorders and respond by providing targeted stimulation. For example, in cases of abnormal

brain activity like seizures or tremors, closed-loop NIs can deliver appropriate stimulation to suppress the undesired activity. Similarly, in the context of treating brain disorders such as stroke, closed-loop NIs can be programmed to provide therapeutic stimulation, aiding in the recovery process [16]. The development of closed-loop NIs holds great promise for personalized and effective interventions in neurological conditions.

In summary, neural interfaces have revolutionized the field of neuroscience by enabling the study of brain activities. Through high-density recording, on-chip processing, and the potential for prosthetic and therapeutic applications, NIs are poised to contribute significantly to our understanding of the brain and pave the way for innovative treatments for brain disorders.

1.3 On-Implant Neural Signal Processing

In most applications (*e.g.*, neuroscience research, brain mapping, and brain-machine interfacing), recording a sufficiently high volume of data is essential in collecting conclusive information from the brain. This is the key drive behind extensive research toward realizing high-density neural recording brain implants. Given strict restrictions in the design and fabrication of brain-implantable devices, recording a huge amount of information and wirelessly transmitting them off the system is not trivial engineering work and faces serious implementation challenges.

1.3.1 Towards High-Density Neural Recording Implants

Figure 1-4 shows the evolution of microfabricated neural recording mircoelectrode arrays in terms of the number of recording sites (a.k.a. *electrodes*) over the past 50+ years. From the introduction of the first microfabricated neural interfacing silicon probe using microelectronic technology in 1969 [17] to mid-2000s, the main engineering focus was on realizing complete systems capable of fully wireless neural recording [4], [5]. With recent reports on the fabrication of microarrays of 5000+ electrodes [8]- [13] the bottleneck in the realization of high-density brain implants has now shifted to 'real-time handling' and 'wireless transmission' of the data they record. The former is a necessity on the application side, and the latter is a technical requirement stemming in the limited bandwidth allocated for wireless interfacing [18] and restrictions in the electric power budget available to a brain-implantable device as mentioned above.

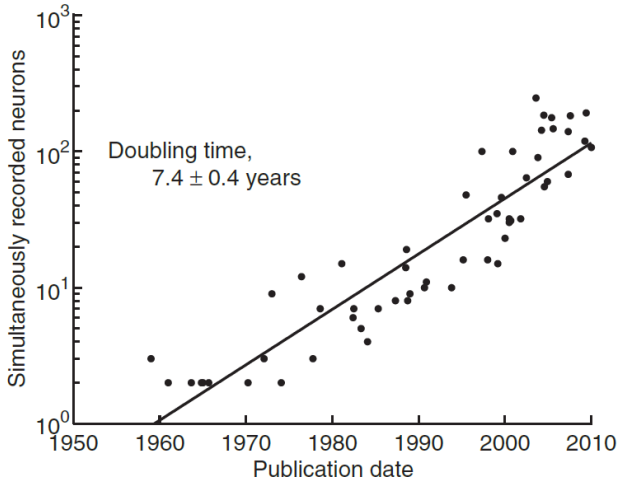


Figure 1-4: Exponential growth in the number of recorded neurons. The number of simultaneously recorded neurons doubled approximately every 7 years [19].

The most effective solution that can confront the aforementioned "*recording density-transmission bandwidth*" dilemma, is the employment of proper signal processing techniques for data reduction/compression. To reduce the extent of the recorded data while preserving the key information in neural signals, there are two general categories of approaches: (a) *data reduction* techniques, in which part of the data conveying no or insignificant information is discarded, and (b) *data compression* techniques, which suggest more compact representation for the recorded data. In neural signal processing, the terms 'data compression' and 'data reduction' are interchangeably used. Moreover, as illustrated in Figure 1-2, signal compression can be performed almost anywhere in the signal path on a neural recording implant. For instance, in the neural

recording front-end, compressive analog-to-digital converters (ADCs) [20], and [21] have been used for signal compression.

1.3.2 Implementation Requirements and Challenges

Physical design and implementation of implantable microsystems faces technical and technological challenges, which can be translated into the following types of considerations:

Physical Considerations- An implant is required to be of the smallest possible size. This is simply because no free space is envisioned inside the body for such a foreign object. Moreover, depending on where it is supposed to be implanted, the device should be of a proper shape and form factor. The size and shape of the building blocks as well as their assembly and integration are all determined accordingly. As a general guideline, all internal circuitry (including on-implant signal processors) should be of the least possible complexity and size, and with preferably no off-chip parts.

Electrical Considerations- An implantable device is energized either using an embedded battery or through wireless power transfer. Either way, the electric power available to an implant is limited, leaving all internal circuitry (including on-implant signal processors) with a serious restriction in the allocated power budget.

Timing Considerations- According to functional requirements in almost all application areas on one hand and the impracticality of envisioning a huge memory capacity for local storage of the recorded data on the other hand, brain-implantable devices are designed to stream neural signals in the real time. As a result, 'on-the-fly' operation is usually a necessity for on-implant signal processing.

1.4 Data Reduction

1.4.1 Truncation

The length of data words throughout signal processing can be reduced by discarding some of the data bits (either unused bits or those of insignificant information) [22], [23], [24]. Referred to as *truncation* (or *quantization*), this action can result in considerable data reduction [22]. Moreover, performing calculations in fixed-point or integer formats rather than floating-point) [25] and implementing scaling with constant factors (*i.e.*, coefficients) using the *canonic signed digit representation* [26] are among the other techniques used at the circuit level to simplify the hardware implementation.

1.4.2 Spike Detection and Extraction

Typical spiking firing rate in an intra-cortical neural signal ranges from 10 to around 150 spike/s. Given that the time course of a typical spike is 1~2ms, a neural signal is, therefore, a sparse signal in which only a small fraction of recorded data carries useful information (*i.e.*, spikes). In applications such as prosthetic devices, only detection of the occurrence of spikes suffices to decode the information of interest [27].

Neural spike detection is usually performed by hard-thresholding, in which a certain attribute of the signal is compared with a given threshold level. The attributes used for this purpose are the amplitude of the signal in the time domain [28] or its absolute value [29], the energy of the signal over a narrow window in time [30], the magnitude of low-pass (approximate) coefficients of the signal in the discrete Haar wavelet space [31], and a measure of the smoothness of amplitude variations [32]. The threshold levels used for spike detection can be defined either manually by the user [33], or automatically by the recording system according to the background noise content of the signal being recorded. In the latter approach, an adaptive threshold level is set about 3~7 times the standard deviation of the background noise above (or below) the baseline level of the signal. Adaptive threshold generation and spike detection have been implemented in both analog and digital domains [15], [28].

In applications where spike waveshapes are required, neural spikes are extracted from the recorded signal. After its occurrence is detected, a spike is extracted in different ways: In [15], only the parts of the spike that goes beyond a pair of adaptive threshold levels (symmetrically defined around the signal baseline or mean value) are extracted. To preserve the integrity of spikes, once a spike is detected, a segment of the signal that accommodates the spike is extracted and telemetered off the implant. The length of this segment can be fixed [34] or adaptively determined according to the length of the spike [31].

1.5 Data Compression

Intra-cortical neural recording stands out from non-invasive recording/imaging techniques like electroencephalography (EEG), Magnetic Resonance Imaging (MRI), and Near-Infrared Spectroscopy (NIRS) due to its superior temporal and spatial resolutions. In the mid-2000s, fully-implantable neural interfacing devices achieved a recording capacity of tens of concurrent channels, making the need for efficient on-implant data compression/reduction inevitable [19]. This need arose from limitations in wireless transmission bandwidth and power allocation for transmitting recorded neuronal activities off the implant. In scenarios where action potentials are considered the primary information-carrying signals, extensive data reduction is accomplished by detecting and extracting neural spikes (also known as action potentials) while discarding

background noise in between [15]. To further enhance data compression rates while preserving individual spike waveforms, a wide range of temporal and spatial compression techniques have been proposed.

From a system-level perspective, two major approaches are taken to compress neural signals on brain implants: hardware-embedded signal compression, and on-hardware signal compression. The former refers to the realization of signal compression in a circuit block that is primarily in charge of a different task in the system (e.g., analog-to-digital conversion [20], [21], data framing, and baseband modulation [35]). The latter is the class of signal compression techniques, for the realization of which a special-purpose processor is designed. Most of the signal compression approaches categorized under this class are based on mathematical transforms. From a signal-level viewpoint, the compression of neural signals can be either temporal or spatial. As illustrated in Figure 1-5, temporal compression is the act of compressing a time series of signal samples recorded on a given channel. On the contrary, spatial compression is applied on the samples (or windows of samples) recorded on multiple channels concurrently at a given instant of time (or over a given time window).

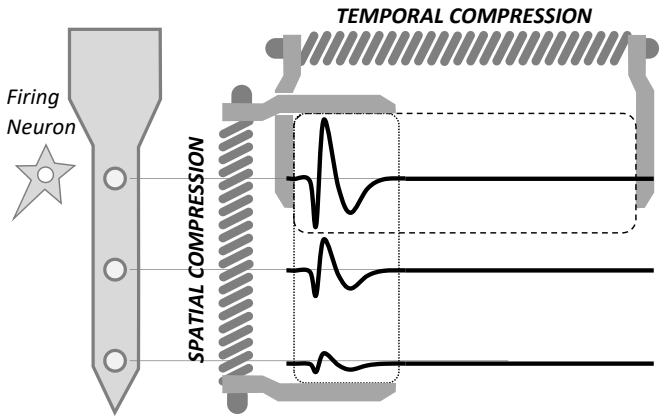


Figure 1-5: Compression of neural signals, temporal vs. spatial [6].

1.5.1 Temporal Compression

Temporal compression techniques generally reduce the data volume associated with consecutive samples within a given time window on a specific recording channel. As an example of hardware-embedded temporal neural signal compression, one can point to the anti-logarithmic ADC reported in [20]. With a physical resolution of 8 bits at the output, this ADC digitizes the background noise with a resolution of 3 bits and large neural spikes with a resolution of up to 10.6 bits. The fact that action potentials (spikes) are sparse events in a neural signal makes this ADC significantly data and power efficient. Similarly, the so-called “delta approach” in [5] is applied to the whole neural signal and achieves lossless data reduction by reporting only the amplitude difference between every consecutive samples rather than their actual amplitude values. On the

contrary, higher data reduction efficiency is achieved by first discarding inter-spike background noise and then applying signal processing techniques on the extracted spikes. This is hereby referred to as *spike compression.* Spike compression on brain implants is usually realized using mathematical transforms. It is important to note that the data reduction achieved in transform-based temporal spike compression techniques is not necessarily the role of the transformation used. Mathematical transforms usually help enhance the discrimination between neural spikes and the background noise [2]. As discussed in [23], the majority of the energy associated with the spikes in a neural signal is conveyed by a limited number of coefficients. As a result, significant data reduction is achieved in the transform domain by discarding the coefficients of insignificant energy (usually by hard-thresholding).

On-hardware, temporal, transform-based compression of neural spikes goes back to 2006, when Oweiss, *et al*, suggested a digital neural signal compressor based on the discrete wavelet transform (DWT) with symlet4 basis function [25]. The existence of digital multipliers in the digital implementation of this transform leads to large chip area and high power consumption, neither of which is in favour of on-implant signal processing. To realize hardware-efficient transform-based spike compression yet achieve comparable signal processing performance, the discrete Haar wavelet transform (DHWT) [31], the Walsh-Hadamard transform (WHT) [36], and the discrete cosine transform (DCT) [37] are subsequently proposed, all implementable with no

need for digital multipliers. To achieve highly efficient data compression, a truncated approximate Tchebichef transform is recently proposed in [38], [39] which is optimized for minimum energy leakage between transform coefficients and is implemented using additions, subtractions, and scaling by small integers (±1, ±2, and ±3).

Principal Component Analysis (PCA) uses tailored transforms to compress data. [40] proposes a stepwise expectation-maximization PCA (sEM-PCA) to reduce the hardware cost. This approach is a lossy approach with reconstruction error up to 8% for a 32-channel recorded data.

In [41], a single-channel ultra-low-power hardware accelerator is proposed for adaptive lossless neural signal compression. This hardware accelerator includes a modified second-order differential pulse code modulator (DPCM) and an adaptive Golomb coding algorithm. The neural signal is first decorrelated and condensed around zero using the DPCM method. Then, the adaptive Golomb coding algorithm compresses the data based on optimal parameters obtained from the signals in real-time. Simulation results show that the average space saving ratio (SSR) is 61.84%. The proposed design is implemented in a 28-nm CMOS technology with an area of 792.4μm^2 and power consumption of 1.05µW at 5MHz.

In [42], [43] a compressed sensing approach is introduced for data compression. Compressed sensing enables to sample a neural signal at a rate lower than the Nyquist frequency without losing

the quality of the signal. However, this approach is not hardware efficient since it needs digital multipliers and large on-chip storage space. The work in [44] presents a simpler representative for compressed sensing by summarizing a multiplications vector. As a result, a more efficient hardware implementation is achieved. However, the power consumption is still too high.

Using Huffman encoding, [45] proposes a band separation technique by first separating the LFP and single spikes and then compressing them independently. Recently, [46] has suggested a lossless approach which is a combination of delta compression and optimized Huffman encoding.

1.5.2 Spatial Compression

Spatial compression techniques involve examining patterns in parallel data streams obtained from multiple channels of simultaneously recorded neural signals. Effective spatial compression methods reported in the literature include compressed sensing [43] and [47], inpainting-based compression [48], modified whitening transform [23], and the MBED technique [49].

The idea of using the whitening transform for reducing redundant activities on neighbouring neural channels is first introduced in [50]. The computationally-heavy calculation of signal correlations (requiring multiplications), however, makes this idea impractical at the time. In 2014, Yazdani, *et al*, propose a *quasi-correlation function*, using which hardware-efficient

implementation of the whitening transform for spatial neural signal compression is made possible [23], [24]. In multi-channel neural recording with high spatial resolution, recorded signals on neighbouring channels share almost the same baseline variations in common. The amplitude of baseline variations is usually much larger than that of neuronal activities. The idea of *Multichannel-Baseline-Extraction Decomposition* suggests that the common baseline is extracted from multiple adjacent channels. Streaming of the extracted baseline along with channel-specific neuronal activities off the implant results in saving a significant amount of bit rate by not reporting redundant baseline data multiple times [49]. Based on this idea, a spatial multi-channel ADC is also developed [21].

As a spatial encoding techniques (realizable in both analog and digital domains) for multi-channel ECoG and LFP signals, [51] generates spatial delta codes in time multiplexed systems (by channel-to-channel subtraction) and achieves 60% compression in both analog and digital domains. Compression in the analog domain helps relax the dynamic range of the analog front-end.

1.6 Spike Sorting

It is a known neuroscientific fact that the extra-cellular action potentials recorded from a given neuron are all of the same general patterns in their amplitude variations [2]. From a signal perspective, however, such spikes are not identical as they are contaminated with random noise, as discussed before. It is interesting that, despite being meaningfully distinguishable from the LFP and the background noise, the extra-cellular spikes recorded from different neurons differ in details such as the number, timing, and amplitudes of the maxima and minima. Therefore, in an intra-cortically-recorded neural signal, the neural spikes of a certain waveshape are associated with the firings of a specific neuron. Figure 1-6 presents multiple occurrences of neural spikes with three different waveshapes, each originated from a specific neuron. These spikes are all extracted from the same neural signal acquired from an *in-vivo* intra-cortical recording [52].

In neural recording for applications such as neuroscience research, brain mapping, and brain-machine interfacing gaining knowledge about the origin of recorded activities is sometimes a necessity. *Spike sorting* is the act of clustering the recorded spikes according to the neuron-specific pattern that exists in their waveshapes [53], [54], [55], [56], [57]. After an initial training (supervised or unsupervised), upon receiving a detected spike, a spike sorter identifies the unit (firing neuron) that has issued that spike. Depending on being supervised or unsupervised, the

process of sorting the spikes is referred to as *classification* or *clustering*, respectively. Reporting the 'label' of the associated 'class' takes significantly less data bits than sending all samples of the spike off the implant. In spike sorting, neural spikes are first detected and aligned. Then, informative features of the spikes are extracted, using which the sorting of the spikes is performed. The online sorting algorithm based on the L1-norm distance [58], spike sorting based on the oblique decision tree [50], spike sorting based on salient feature selection and window discrimination [3], and correlation-coefficient-based spike sorting [59] are examples of hardware-efficient, online, on-implant spike sorting algorithms reported in the literature. It is important to highlight that the salient feature selection employed in [3] is specifically tailored for spike sorting techniques and differs significantly from the salient sample selection utilized in the current study. As we will illustrate further, the salient sample selection introduced in this study serves the dual purpose of downsampling and segmentation of a spike.

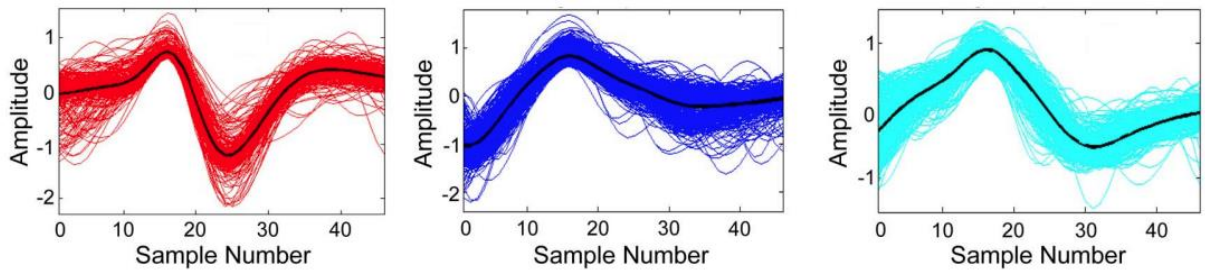


Figure 1-6: Three separate classes of spikes with totally different waveshapes [52].

1.7 Quantitative Measures

In on-implant signal processing, the merit of tasks and operations is not only in the processing itself, it also be determined in how well it is implemented in hardware. As such, the success and merit of such tasks are assessed using *signal processing measures* and *hardware measures.* Moreover, given the nature of the signals being processed, variabilities such as spike classes and random noise contamination need to be taken into consideration when assessing the performance of signal processing on brain-implantable devices.

1.7.1 Signal Processing Measures (Performance Measures)

Depending on the category of signals processing (discussed in previous sections), the success and quality of the processing are assessed using specific performance measures, some of the most common of which are introduced as follows:

Spike Detection Accuracy- The accuracy of detecting spikes in a neural signal is determined by taking into consideration the number of actual spikes that are correctly detected (known as *True Positives-TP*) as well as the number of non-spike signal fluctuations that are incorrectly counted as spikes (referred to as *False Positives-FP*). *Spike detection accuracy* (*SD.Acc.*) is defined as:

$$SD.Acc.\% = \frac{TP - FP}{Total\ No.\ of\ Spikes} \times 100\% \tag{1.1}$$

Compression Rate- In neural signal compression, the extent of reduction in the data representing the recorded signals referenced to the raw digital data acquired by the recording front-end is quantified as the *compression rate* (*CR*):

$$CR = \frac{\#\ of\ Bits\ in\ the\ Raw\ Signal}{\#\ of\ Bits\ in\ the\ Compressed\ Signal} \tag{1.2}$$

As mentioned before, in a wide spectrum of neural signal compression techniques, a great deal of data compression is achieved through spike extraction, and a portion of it comes from data reduction techniques such as spike compression, truncation, and compact data representation.

True Compression Rate- The portion of the compression rate that is achieved through spike extraction proportionally depends on the *firing rate* of spikes in the neural signal being compressed. This evidently makes the CR a function of both the signal compression technique used and the firing rate of spikes. To be independent of the spike firing rate, the *true compression rate* (*TCR*) is introduced in [31] as:

$$TCR = CR \times \frac{Spike\ Firing\ Rate}{1\frac{Spike}{Sec}.} \tag{1.3}$$

This is a measure that can be taken as either the CR normalized to the spike firing rate of the signal, or the CR for a firing rate of 1 spike per second.

Reconstruction Error- The accuracy of curve fitting is measured by calculating the normalized root-mean-square value of error normalized to the peak-to-peak amplitude of the spike (Normalized RMS of Error (NRMSE)) expressed as:

$$NRMSE\% = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(R_i - S_i)^2}}{A_{P-P}} \times 100 \tag{1.4}$$

where R_i is the i^{th} sample of the reconstructed spike, S_i is the i^{th} sample of the original spike, N is the number of samples in each spike, and $A_{P\text{-}P}$ is the peak-to-peak amplitude of the spike under study. The acceptable reconstruction error in neural signal compression techniques is contingent upon the particular requirements and constraints of the application. Nevertheless, neural scientists and engineers generally find a reconstruction error below 10% satisfactory, as it indicates a reasonably close match between the reconstructed spike and the original spike [2].

Signal-to-Noise Ratio- As mentioned before, neural spikes are mainly taken as the signal in the majority of brain-implant applications. The strength of the spike peak-to-peak amplitude with respect to the noise amplitude is, therefore, introduced as the *signal-to-noise ratio* (*SNR*). Even though there are different definitions for SNR in this field, perhaps the most common way of measuring the SNR is the ratio of the spike peak-to-peak amplitude to the double standard deviation of the background noise [24]:

$$SNR = \frac{Spike\ peak-to-peak\ amplitude}{2 \times Standard\ deviation\ of\ noise} \tag{1.5}$$

Sometimes the SNR is expressed in decibels [24]:

$$SNR_{dB} = 20log_{10}(\frac{Spike\ peak-to-peak\ amplitude}{2\times Standard\ deviation\ of\ noise}) \quad (1.6)$$

From an electrophysiological perspective, a good neural signal has a typical SNR of ~10 or higher.

1.7.2 Hardware Implementation Measures (Circuit-Level Measures)

Hardware implementation measurements refer to the quantifiable characteristics or parameters that are assessed during the design, fabrication, testing, and evaluation of hardware systems. These measurements provide important information about the performance, quality, and reliability of the hardware implementation. Here are some common hardware implementation measurements:

Power Consumption- Power consumption measurement is crucial to determine the energy efficiency of hardware systems. It involves quantifying the amount of electrical power consumed by the hardware components during operation, both in active and idle states. Lower power consumption is desirable for improving battery life in portable devices or reducing energy costs in large-scale systems.

Clock Frequency- Clock frequency measurement indicates the speed at which a hardware system operates. It represents the number of clock cycles per unit of time, typically expressed in kilohertz (kHz) megahertz (MHz). Higher clock frequencies generally result in faster computational performance, however; they can also impact power consumption and heat dissipation.

Throughput- Throughput measurement quantifies the rate at which data can be processed or transmitted by a hardware system. It represents the amount of data or operations that can be completed within a given time period. Higher throughput indicates improved system performance and efficiency, especially in applications that require high-speed data processing or communication.

Latency- Latency measurement refers to the time delay between initiating an action and observing a response or output from a hardware system. It is particularly important in real-time applications where low latency is crucial for maintaining responsiveness. Latency can be influenced by various factors, including processing time, communication delays, and memory access times.

Heat dissipation- measurement assesses the ability of a hardware system to manage and dissipate heat generated during operation. Excessive heat can degrade system performance, cause

component failure, or pose safety risks. Efficient heat dissipation mechanisms, such as heatsinks or fans, are employed to maintain optimal operating temperatures. These hardware implementation measurements help engineers and designers assess the performance, efficiency, quality, and reliability of hardware systems, enabling optimization, troubleshooting, and improvement in various technological domains.

1.8 Thesis Overview

This thesis proposes a novel neural spike compression technique dedicated to high-density brain implants. Chapter 1 reviews signals of interest in neural engineering study, on-implant neural signal processing, data reduction, and data compression techniques reported in literatures. Chapter 2 is dedicated to illustrate a data reduction framework, specific to the compression of extra-cellular neuronal action potentials on brain-implantable microsystems. The proposed approach significantly minimizes the amount of data needed to represent spike waveforms, facilitating the development of next-generation, high-density neural recording brain implants. Employing a highly compressive strategy, the framework selects a limited number of essential samples from the spike. These samples, along with predefined functions, are used to construct the entire spike waveshape. The amplitudes and timings of these key samples are transmitted from the implant to reconstruct the spike waveshape on the external side of the system. Beyond its exceptional data compression

capabilities, this technique is highly efficient in terms of hardware, making it well-suited for brain-implantable neural recording microsystems with a high number of channels. With an average spike firing rate of 8 spikes/s, the circuit achieves temporal reduction of neural data at an average compression rate of approximately 272, equivalent to a true compression rate of around 2176. Based on the proposed technique, in order to have a better understanding of the technique, a comprehensive noise analysis is elaborated in chapter 3. This chapter studies two major effects of the existence of noise. At the end, the proposed technique in the presence of both effects is investigated and two innovative metrics for the assessment of the quality of the signals (neural spikes) before and after signal processing are introduced. Chapter 4 presents the hardware design and implementation of the proposed idea, along with the corresponding experimental results. Finally, future works and conclusions are presented in chapter 5.

1.9 Conclusions

It is common in neural signal compression that a significant extent of data reduction is achieved through spike extraction [15], and some extra reduction is resulted from temporal spike compression (e.g., [22], [36], [37]). Table 1-1 showcases a detailed performance comparison between the currently available temporal and spatial compression techniques. The objective of compression techniques is to reduce the size of data while preserving its essential information,

thereby optimizing storage and transmission resources. Temporal compression primarily focuses on eliminating redundancy within sequential data over time. By identifying and representing patterns or changes that occur across consecutive frames or timestamps, temporal compression can significantly reduce data size without compromising the overall content. On the other hand, spatial compression concentrates on removing redundancy within individual frames or spatial data. This process involves analyzing and encoding the spatial structure and relationships between various data points within a single frame, leading to a more compact representation.

Table 1-1: Comparison of Compression Rates between the existing techniques

Math. Technique		**DWT/ Symlet4**	**WHT**	**DWT/ Haar**	**Modified Whitening Transform**	**MBED**	**DTT**
Reference		[25]	[36]	[22]	[24]	[49]	[39]
Year		2007	2014	2015	2018	2020	2023
Technique Type		Temporal	Temporal	Temporal	Spatial	Spatial	Temporal
Technology (µm)		0.5	0.18	0.13	0.18	0.13	NA
Supply Voltage		3.3	1.8	1.2	1.8	1.2	NA
Sampling Rate (kS/sec.)		25	25	20	20	NA	20
No. of Channels		32	128	64	32	16	256
Firing Rate		NA	8	8	10-100	NA	35
Compression Measure	CR	63	62	116	13	48	26.15
	TCR	504	496	903	130-1300	NA	915
Area (mm^2)		5.7	1.64	0.206	0.288	0.048	2.88
Area/Ch. (mm^2)		0.178	0.0128	0.0032	0.009	0.003	0.0112
Power (µW)		3008	83.2	94.08	238	102.4	NA
Power/Ch. (µW)		94	0.65	1.47	7.43	6.4	NA

Chapter 2 Selective Downsampling: An Innovative Idea in Neural Spike Compression

2.1 Introduction

This chapter is devoted to the introduction of the novel implant-appropriate spike compression technique proposed in this thesis. We first have a glance at the motivation in this work and the challenges we face in the realization of the proposed idea. Then, the details of the proposed idea will be presented, and the idea is verified by simulation using a dataset that is composed using 8 different real spike waveshapes acquired from *in-vivo*, extra-cellular recordings.

2.2 Motivation and Challenges

What is introduced in this thesis is a novel approach for achieving temporal data reduction in brain implants, which is based on spike compression. The primary objective of this approach is to address the pressing need for enhancing the rate of data compression in such implants. By leveraging spike compression, we aim to efficiently reduce the amount of data generated by

implantable devices. An additional benefit of this proposed technique is spike denoising, which stems in the curve fitting nature of this approach.

This study presents a novel framework for data reduction specifically tailored to extra-cellular neuronal action potentials. The proposed framework effectively minimizes the amount of data required to represent spike waveforms, thereby enabling the development of next-generation, high-density neural recording brain implants. By employing a highly-compressive approach, this framework selects a limited number of significant samples from the spike, and utilizes them to interpolate the entire waveform. The amplitudes and timings of these salient samples are transmitted from the implant to reconstruct the complete spike waveform externally. In addition to its exceptional data compression capabilities, this technique boasts remarkable hardware efficiency, making it ideal for brain-implantable neural recording microsystems with numerous channels.

2.3 Proposed Idea

As shown in Figure 2-1, it is assumed that first, the neural recording system under study comprises an implantable module and an external module, which are linked through wireless connection. Second, pre-detected spikes from a data set is applied to the proposed compressor's

input. In the proposed method, the pre-detected spike is represented using the start and end points as well as global and local extremum points of its waveshape, which are all referred to as salient samples, hereafter. For a given spike, first, the salient samples are extracted on the implant module, which can be taken as a 'selective downsampling' step. Then, the key attributes of the salient samples are transferred to the external module. On the external side, some predefined functions are fit to the received salient samples in order to reconstruct the spike waveshape. Through sending the timing (sample index) and amplitudes of only the salient samples (rather than telemetering the amplitudes of all the spike samples) from the implant to the external module, significant data reduction is achieved.

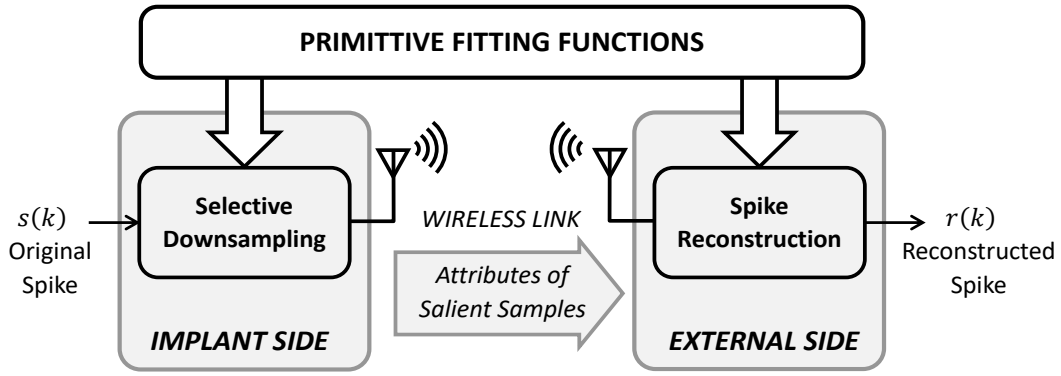


Figure 2-1: Functional diagram for the implementation of the proposed data reduction approach on a high-density brain implant.

The proposed method offers three main advantages: (a) the volume of the data transmitted off the implant is significantly reduced, (b) the reconstruction of the spike waveshapes using

predefined smooth curves allows for the removal of amplitude noise contamination of the recorded spikes, and (c) the low complexity of the computations on the implant side, and shifting the major part of the computations to the external side are the key aspects of the proposed idea, which make it a proper candidate for hardware-efficient, on-implant, neural signal compression.

2.3.1 Salient Samples

In the context of analysing neural spikes, identifying salient samples is crucial for understanding the characteristics of the spike waveform. These salient samples typically include the start and end points of the spike, as well as the samples that correspond to the highest and lowest points of the waveform. To pinpoint the extremum points accurately, a method is employed that involves detecting changes in the slope of the spike. This means examining the rate of change between consecutive samples in the spike waveform. When the slope of the waveform changes from positive to negative or vice versa, it signifies the presence of an extremum point. By detecting these significant samples within the neural spike, researchers and scientists can gain valuable insights into the shape, duration, and overall characteristics of the spike waveform. These salient samples are often used for further analysis and processing in various applications, such as spike sorting, spike detection, and feature extraction, among others.

2.3.2 Sample Triplets

To identify the extremum points with rather low computational complexity, every three consecutive samples of the spike under study is taken as a *sample triplet* as illustrated in Figure 2-2. Each sample triplet is represented by a *triplet representative sample* (*TRS*). The amplitude of the TRS is basically the average of the triplet samples:

$$trs(m)=[s(3m) + s(3m+1) + s(3m+2)]/3 \tag{2.1}$$

in which $s(i)$ is the i-th sample of the neural spike, and m is the index of both the *sample triplet* under study and the associated TRS amplitude. A change in the sign of the slope of the TRS-spike at index m is an indication of the existence of an extremum point there. Comparison of the amplitudes of the three samples in that triplet determines which one is the extremum point. When $trs(m)-trs(m-1)>0$ and $trs(m+1)-trs(m)<0$, the m-th sample triplet contains a maximum point, and that is the sample with the largest amplitude in that triplet. Otherwise, a change of slope sign from negative to positive indicates that there is a minimum point in the m-th sample triplet, which is the sample with the smallest amplitude in that triplet. It is worth noting that, to further simplify the computational complexity of finding spike extremum points, the divide-by-3 is omitted from (2.1) as it has no effect on locating extremum points. It should be noted that when averaging two consecutive samples of the spike under study, there are not two slopes available to identify the

extremum point. Furthermore, extending beyond three consecutive samples necessitates additional hardware resources.

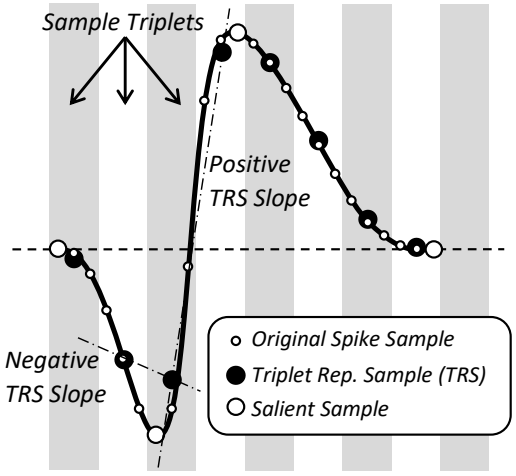


Figure 2-2: Sample triplets, triplet representative samples (TRS), and TRS slopes on a typical spike waveshape.

2.3.3 Downsampling and Segmentation

In general, when downsampling a signal that is already sampled according to the Nyquist criterion [60], the preservation of the key signal information becomes a critical concern. In the idea proposed in this work, part of the signal information is conveyed by the fitting function(s) used for spike waveshape reconstruction. If properly chosen, the curvature of the fitted reconstructive functions will retrieve part of the inter-sample variations that are lost as a result of downsampling.

There are two key issues that contribute to defining the details of the proposed spike compression approach: (a) The typical spike waveshape is too complicated to be modelled using ordinary algebraic or trigonometric functions, and (b) spike waveshapes vary in so many morphological aspects from one neuron to another. Therefore, as illustrated in Figure 2-3, spike waveshapes are broken into multiple primitive monotonic segments, P_iP_j, with the junction points, P_i and P_j, being salient samples. Attributes of the salient samples are then framed and transmitted to the external side of the system. On the external side, each segment is fitted by a smooth curve, which is formulated using a primitive fitting function, y_{ij}. Finally, the waveshape of the received spike is reconstructed by concatenating the retrieved spike segments.

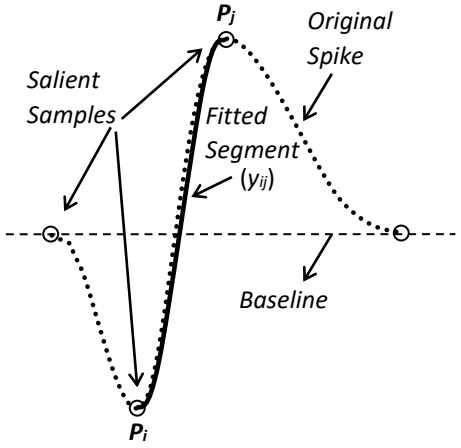


Figure 2-3: Illustration of the proposed idea, Waveshape of a typical neural spike (dotted line), the concept of salient samples (circles), and the fitting function used to model the segment waveshape (solid line)

2.3.4 Primitive Fitting Functions

The fitting functions used in the proposed approach are chosen from a library of polynomial functions. A first-order polynomial function is unsuitable for interpolating the curvature of a spike as it lacks the necessary complexity for capturing intricate shapes. Moreover, it's worth noting that a second-order polynomial is incapable of producing a line with a zero slope at both ends. As a result, in this work, it is suggested that each segment is interpolated primarily using a third-degree polynomial function expressed as:

$$y(kT) = C_3(kT)^3 + C_2(kT)^2 + C_1(kT) + C_0, \tag{2.2}$$

where k is the sample index and T is the sampling time of the digitized signal. Using a fourth-order polynomial for interpolating a segment requires an additional sample for more precise segment reconstruction. However, employing polynomial functions beyond the fourth order degrades the compression rate and needs more resources to be implemented. To simplify the discrete-time equations, as it is common, the sampling time is taken equal to 1 (T=1) hereafter, and the signals are represented in terms of the sample index k. The coefficients $C_0 \sim C_3$ in (2.2) can, therefore, be determined according to the attributes of the start and end samples of the segment, P_i (i , A_i) and P_j (j , A_j), as well as the slopes of the signal at those samples. It is worth mentioning that, to simplify calculations on the implant side, instead of sending the aforementioned slopes, what is reported to

the external setup is the 'forward difference' between the amplitudes of the start sample and the sample next to it (*i.e.,* $\Delta_{F,i} = A_{i+1} - A_i$), and the 'backward difference' between the end sample and the sample prior to it (*i.e.*, $\Delta_{B,j} = A_j - A_{j-1}$).

2.3.5 Slope Considerations

Knowing that a typical spike has a continuous and derivable waveform, it can be said that the slope of spikes at extremum points is always zero. Assuming that a spike has a gentle rise from the baseline level at the beginning and also gradually approaches the baseline in the end, the slopes

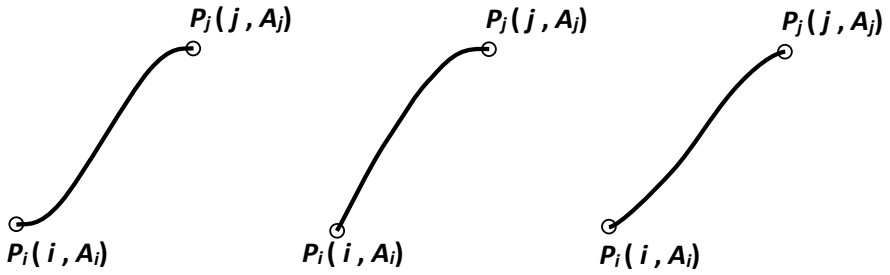


Figure 2-4: Some of the waveshape types (with and without zero slopes at either end) that can be generated using a third-degree fitting function.

at both ends of spikes can be taken equal to zero as well. This means that the slopes at the start and end of each and every segment of a typical spike are always equal to zero. In this situation (shown as the left-most segment in Figure 2-4) the third-degree fitting function is simplified to:

$$y(kT) = C_3(kT)^3 + C_2(kT)^2, \tag{2.3}$$

in which

$$C_3 = \frac{2A_{ji}}{(\Delta T)^3} \;\; \& \;\; C_2 = -\frac{3A_{ji}}{(\Delta T)^2} \tag{2.4}$$

where $A_{ij} = A_j - A_i$ and $\Delta = j - i$. However, even though the slopes at extremum points are always 'mathematically' equal to zero in continuous time, this might not hold true in the discrete time (*i.e.*, for a sampled spike). The situations where the slope at an extremum point in discrete time can or cannot be taken equal to zero are shown in Figure 2-5. Specifically speaking, when a given segment, P_iP_j, spans a rather large amplitude during a short period of time, (the absolute value of) the slope of the spike when going from the segment start sample to the sample immediately next to it (*i.e.*, from P_i to P_{i+1}) or from the segment end sample back to the sample immediately prior to it (*i.e.*, from P_j back to P_{j-1}) is likely to be considerably greater than zero. The criterion for a 'large segment amplitude span' is when it is greater than half the full-scale amplitude range, V_{FS} (*i.e.*, $|A_{ji}|>V_{FS}/2$). In this case, in order for the suggested fitting function to fit such a spike segment more accurately especially at both ends, the original 4-term polynomial of (2.2) is used in full.

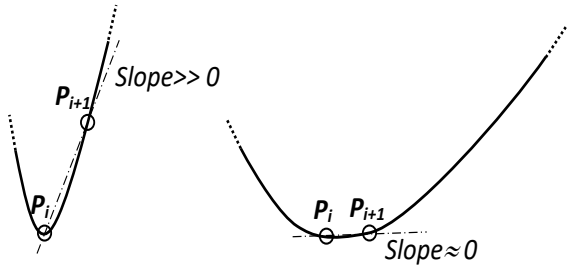


Figure 2-5: An extremum point (P_i) at which the slope cannot (left) or can (right) be approximated to zero

2.3.6 Rivet Samples

The curve fitted to a spike segment lies exactly on the original segment at both ends (*i.e.*, at P_i and P_j), and might be to some extent off elsewhere. As a result, the reconstruction error is equal to zero at both ends of spike segments and non-zero otherwise. Even though for most intra-cortical neural spikes with typical waveshapes the 3rd-order polynomials of (2.2) or (2.3) fit all spike segments sufficiently accurately, there are still exceptions where the reconstruction error is unacceptably high. As a remedy, as illustrated in Figure 2-6, the fitted curve is *rivetted* to the original spike segment at a sample in the middle of the segment, almost where the fitting error peaks. The fitting error is, consequently, zeroed at the *rivet sample*, P_R, and the maximum error is therefore significantly lowered. In this situation, to allow for a more accurate curve fitting, a five-term 4th-order polynomial function is used:

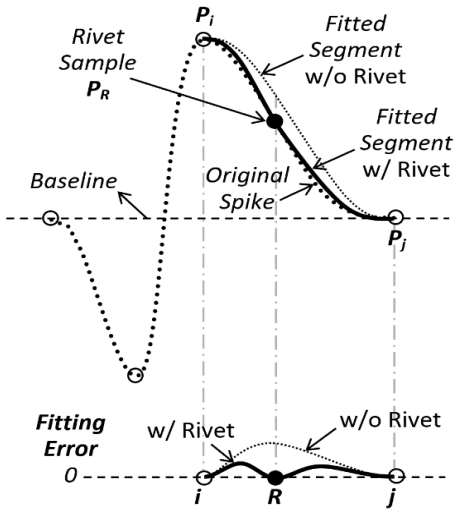


Figure 2-6: Adding a 'rivet sample' to a segment to significantly enhance fitting accuracy

$$y(kT) = C_4(kT)^4 + C_3(kT)^3 + C_2(kT)^2 + C_1(kT) + C_0, \tag{2.5}$$

The coefficients C_4-C_0 in (2.5) are calculated on the external side of the system using the attributes of the segment start and end samples and the rivet sample. The location of the rivet sample cannot be determined on the implant side according to where the fitting error is maximized as it would require doing the segment reconstruction computations on the implant side. Therefore, always the 6th sample after the segment start sample is experientially taken as the rivet sample. This way, there will be no need for reporting the timing of rivet samples.

2.3.7 Data Set to Verify the Proposed Idea

To verify the efficacy and evaluate the performance of the proposed spike compression technique, a dataset is composed using 8 different real spike waveshapes acquired from *in-vivo*, extra-cellular recordings [61], [62], [63]. To study the impact of noise on the performance of the proposed technique, the SNR spans from 5dB to 25dB for each individual spike waveshape. Figure 2-7 presents samples of the 8 spikes used in this work with a typical duration of 2ms, sampled at of 25 kSample/s. To exhibit the success of the spike compression technique proposed in this work, in the spikes used for the tests, the number of extremum points range from 1 to 6, and the order and amplitudes and also timings of the extremum points are all different.

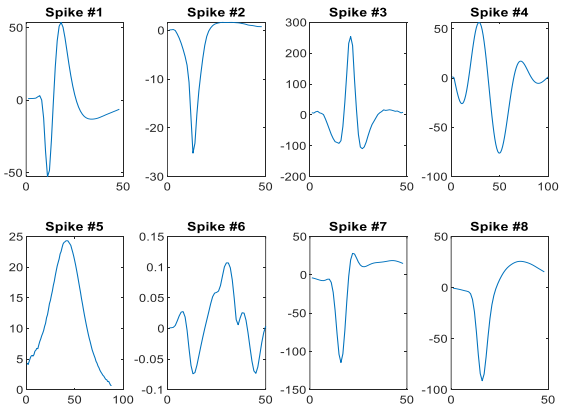
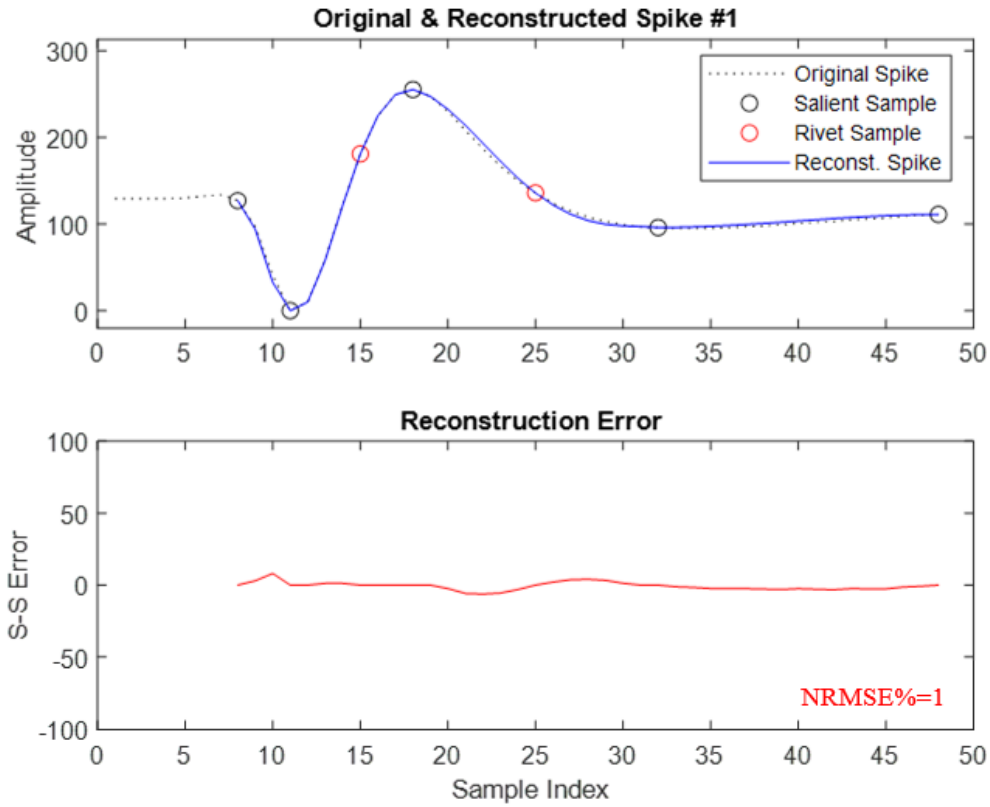


Figure 2-7: Eight different spike waveshapes acquired from in-vivo recording [61], [62], [63]. These spikes are used to verify the functionality and assess the performance of the proposed spike compressor.

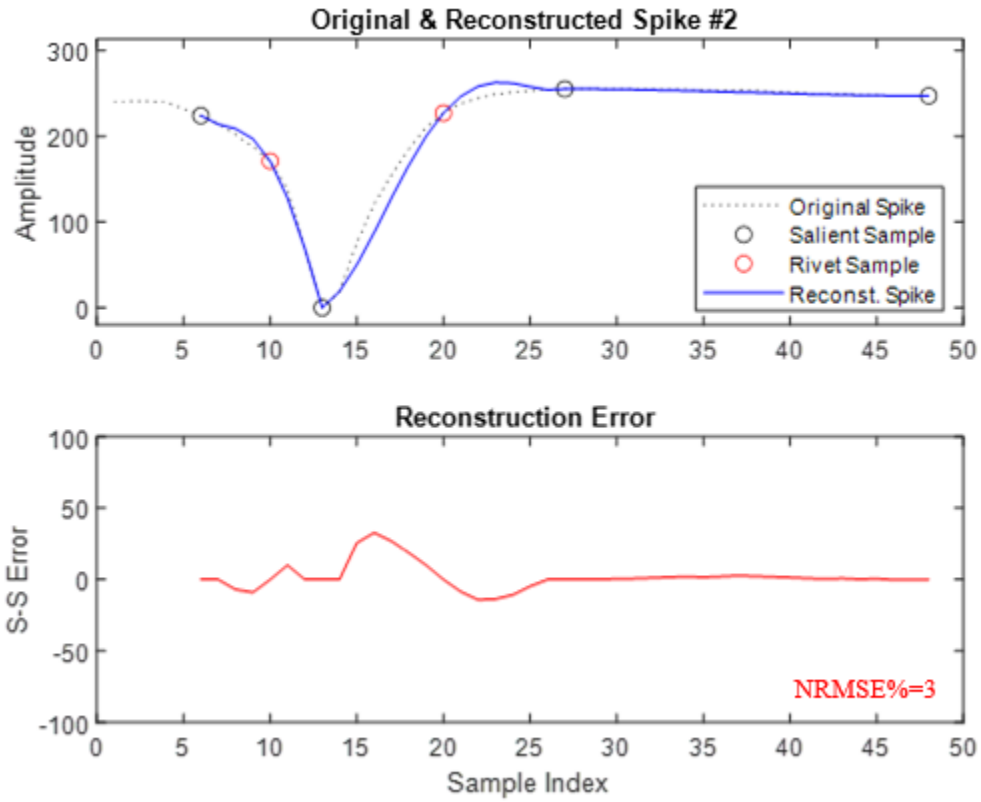
2.3.8 Simulation Results

Figure 2-8 shows the performance of the segmentation and curve fitting for eight sample spikes. In this test, only noiseless spikes are used to report merely the contribution of the curve fitting error in the reconstruction of spikes. In this test, only noiseless spikes are used to report merely the contribution of the curve fitting error in the reconstruction of spikes. As mentioned in Figure 2-8, the maximum NRMSE for the four spikes used in this test is as low as 4%. What is important to note in this test is that the proposed approach works for spikes with 5, 6, and 7 salient samples.



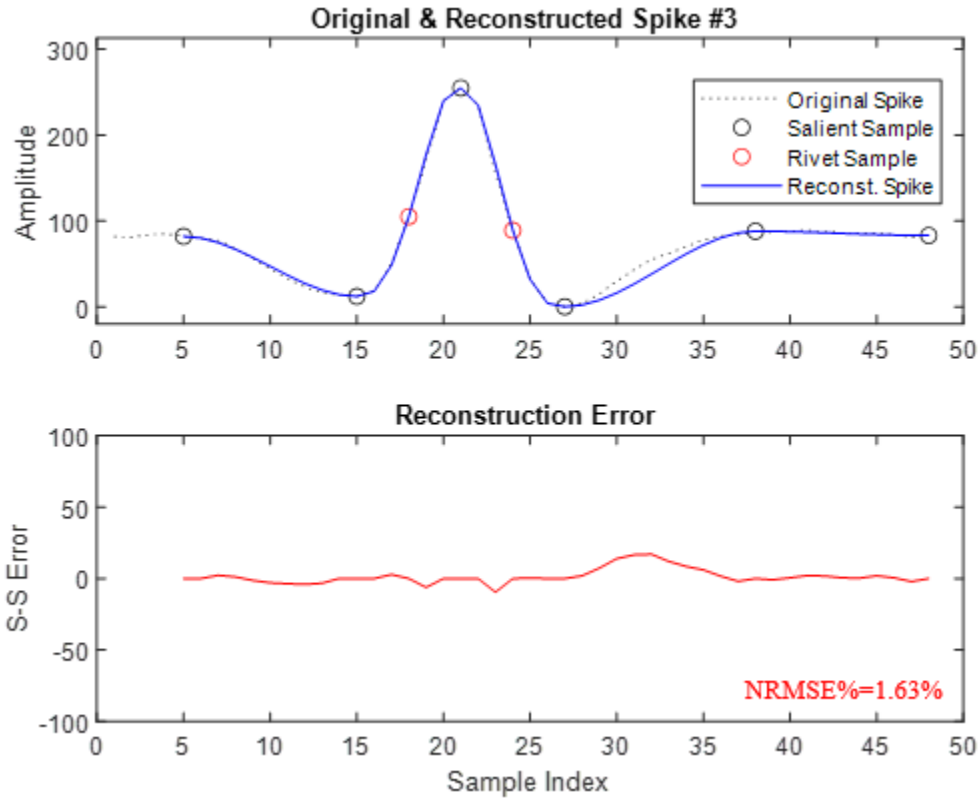
(a)

Figure 2-8 Simulation results demonstrating the proposed technique showing the original spike (dotted line), reconstructed spike (solid line), salient samples (black circles), and rivet samples (red circles) in the upper plot, and the sample-to-sample reconstruction error (and the normalized RMS of error, NRMSE) in the lower plot; (a)-(h) present the results for spikes #1~#8, respectively



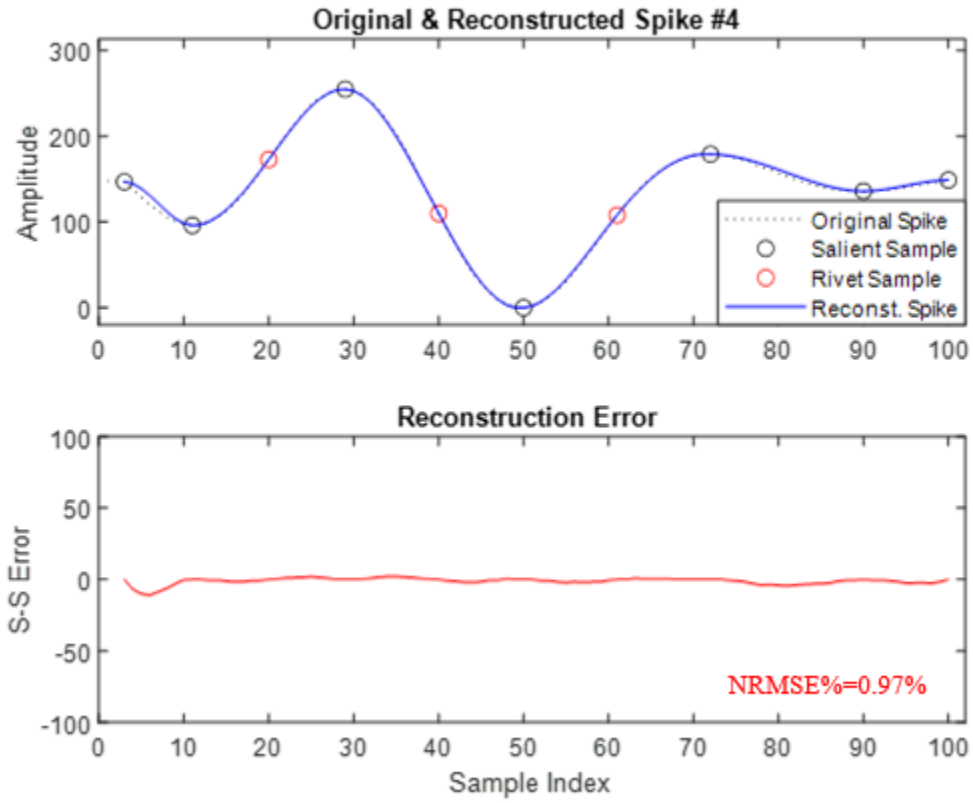
(b)

Figure 2-8 (Continued)



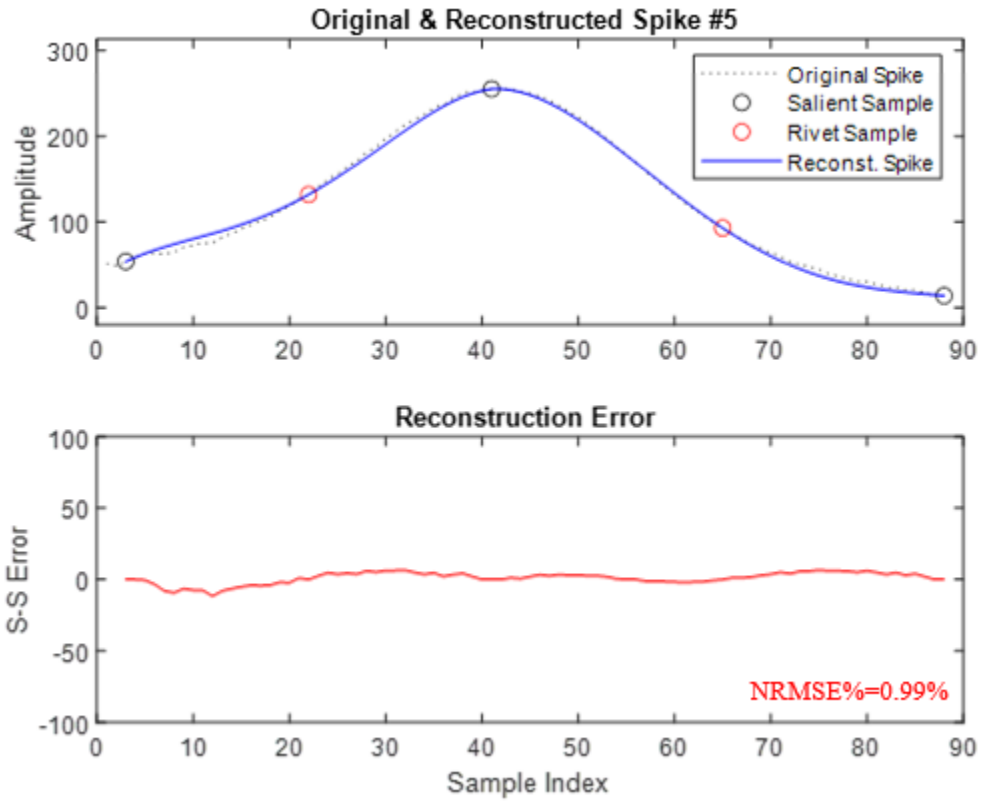
(c)

Figure 2-8 (Continued)



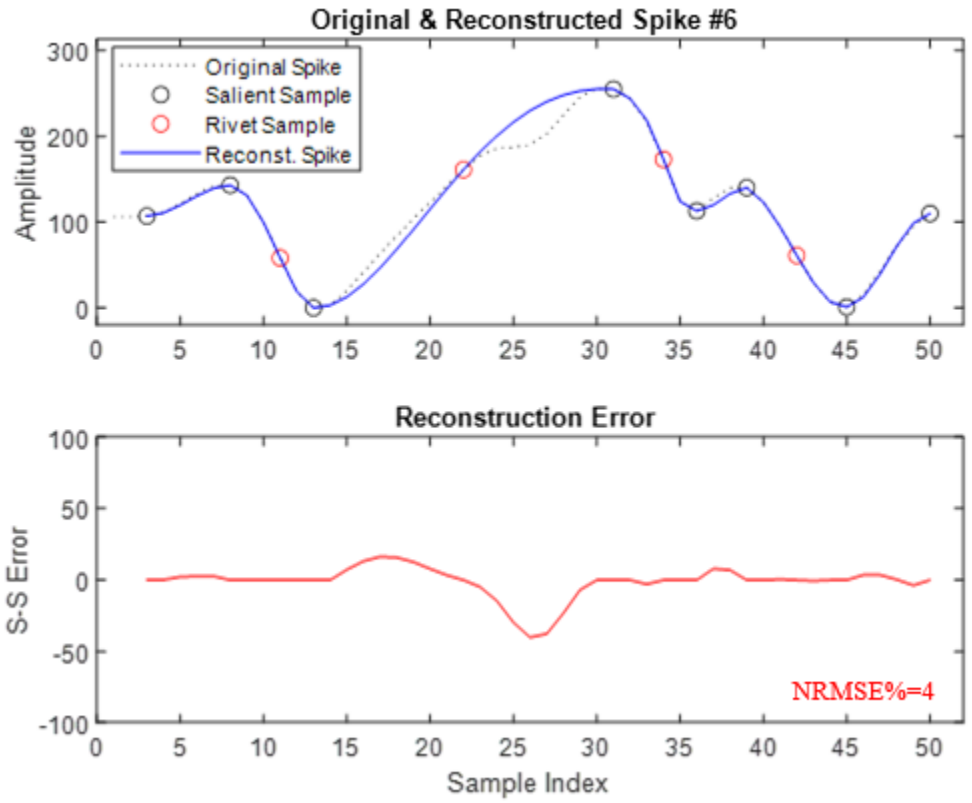
(d)

Figure 2-8 (Continued)



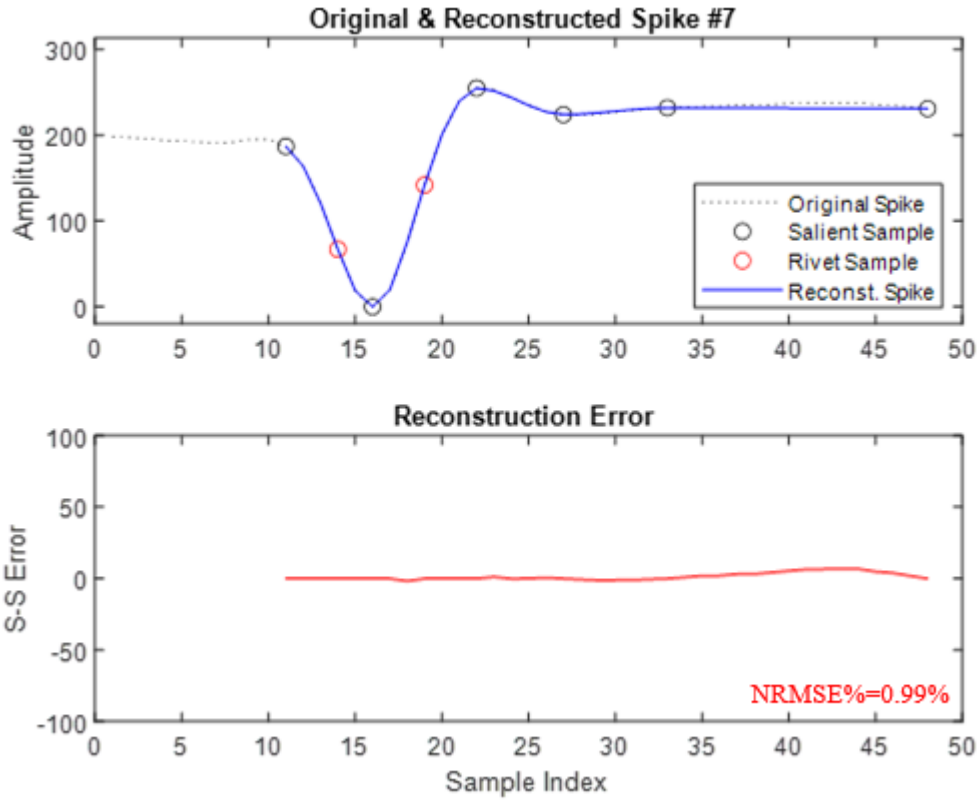
(e)

Figure 2-8 (Continued)



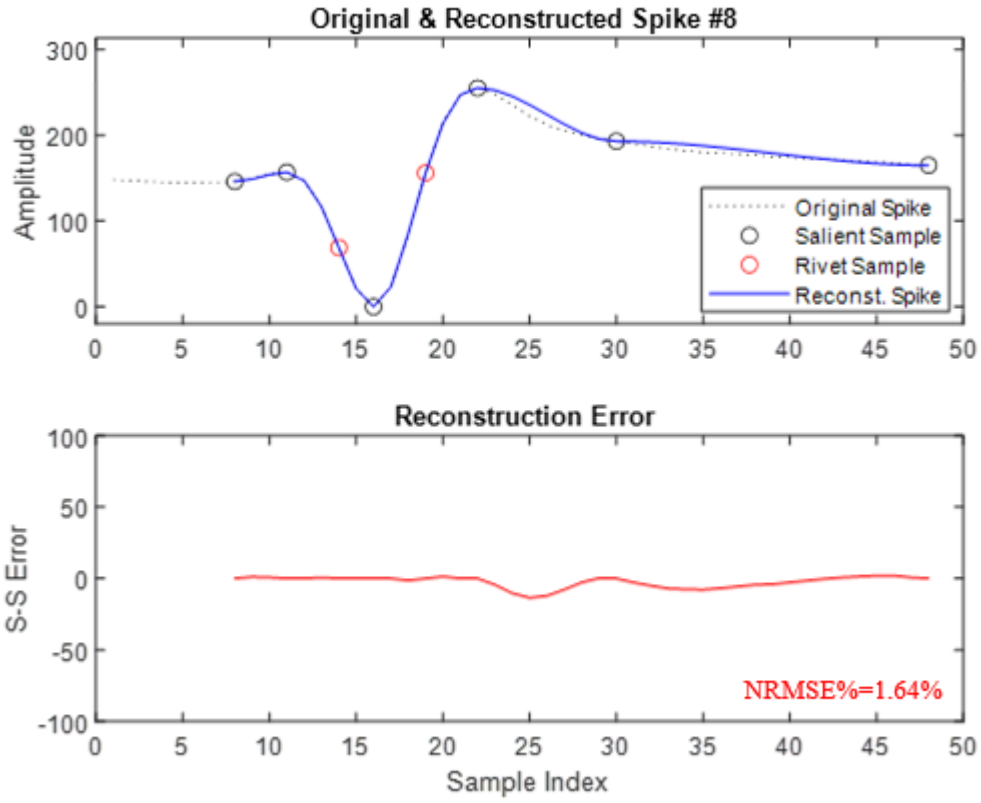
(f)

Figure 2-8 (Continued)



(g)

Figure 2-8 (Continued)



(h)

Figure 2-8 (Continued)

2.3.9 Rule-Based Spike Reconstruction

Once the attributes of salient samples of a given spike arrive at the external side of the system, the associated waveshape of that spike is reconstructed using a rule-based procedure. Figure 2-9 shows the flowchart of the spike reconstruction procedure. The procedure first locates the salient samples according to the amplitudes and timings received from the implant module. It then identifies the number of spike segments, and calculates the associated heights. Based on the height of each spike segment, the degree of the primitive fitting function is decided upon. If the segment height is greater than half the full-scale range, V_{FS}, the 4^{th}-degree polynomial of (2.5) is used as the fitting function; otherwise, the two-term 3^{rd}-degree polynomial given in (2.3) is employed. In the former case, the polynomial coefficients are calculated using the timings, amplitudes, and slopes of the salient samples and the amplitude of the rivet sample accompanying the segment information. In the latter case, coefficients of the polynomial are simply computed using the timings and amplitudes of the segment start and end. Finally, the segments are concatenated and a unified spike is derived.

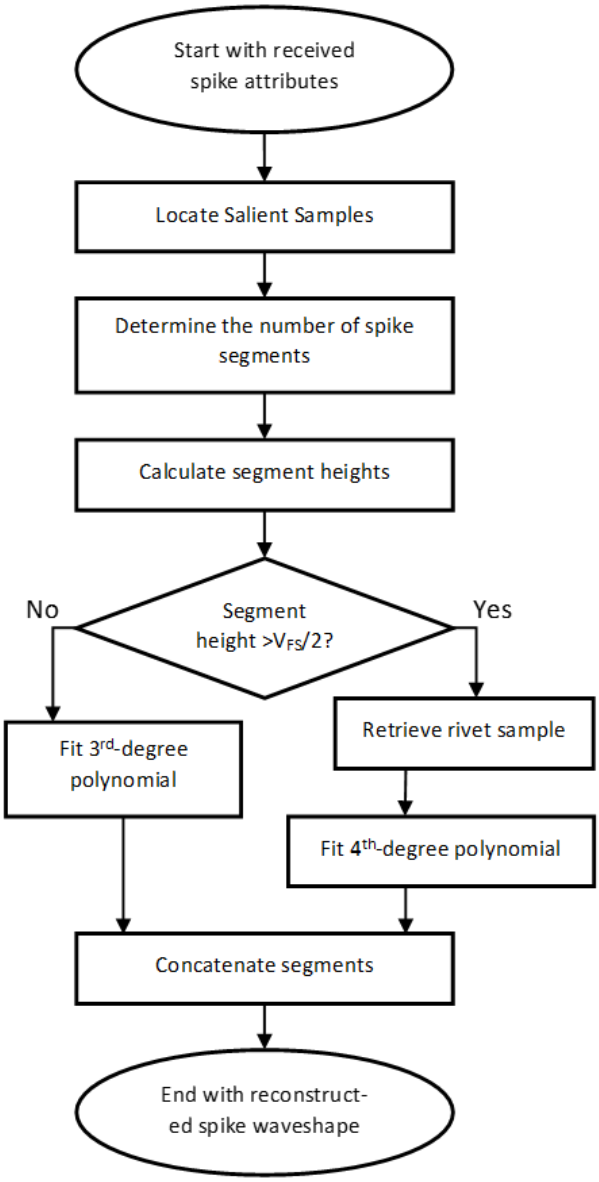


Figure 2-9: Flowchart of the rule-based spike waveshape reconstruction procedure.

2.4 Research Methodology

This research proposes an innovative temporal compression idea based on selective downsampling, in which the amplitudes and timings of some salient samples for each spike are transferred to the external side of the system. To enhance the accuracy of the proposed idea, slope considerations are investigated. As a result, when a given segment spans a rather large amplitude during a short period of time, the slopes at both ends of the segment are considered as another type of salient sample attributes and sent off the implant. Additionally, the concept of *rivet samples* is introduced to minimize reconstruction errors. Then, a data set with eight different waveshapes are used to verify the proposed compression method. First, the study uses noiseless spikes to isolate the impact of curve fitting errors in spike reconstruction. Furthermore, this research examines the influence of noise on the novel technique, highlighting its effects on extremum samples replacements and the fitted function fluctuations. This study, shows that noise might replace the extremum samples and fluctuate the fitted function. These two major effects of noise are quantified and qualified in chapter 4. Finally, this research introduces two signal quality assessment metrics, demonstrating the spike denoising capability of the proposed approach.

2.5 Conclusions

This chapter presents a novel technique for compressing neural signals. The proposed approach revolves around utilizing the salient samples of a typical spike to reconstruct it externally. Consequently, the implantable microsystem is only required to transmit the salient samples and their associated time stamps to the external side. This methodology achieves a noticeable compression rate. Additionally, the chapter introduces several techniques aimed at enhancing the accuracy of the reconstructed spike, albeit at the expense of a smaller compression rate.

Chapter 3 Noise Analysis

3.1 Introduction

In the previous chapter, a data reduction framework, specific to extra-cellular neuronal action potentials was introduced. The proposed framework significantly reduces the extent of data representing spike waveforms, paving the way for the implementation of next-generation, high-density neural recording brain implants. This highly-compressive approach picks a small number of salient samples of the spike, using which and based on some predefined functions the entire spike waveshape is interpolated. The amplitudes and timings of the salient samples are sent off the implant in order to reconstruct the spike waveshape on the external side of the system. As long as the spike under study is noiseless, selective downsampling and subsequently fitting smooth curves to the resulting salient samples works quite well. In reality, however, the non-negligible amplitude noise that is unavoidably mixed with the spike inevitably affects the idea proposed in this work. As shown in Figure 3-1, this chapter studies two major effects of the existence of noise: (1) extremum point displacement, and (2) the impact of noise-related salient sample amplitude fluctuations on the fitted functions. At the end, the proposed technique in the presence of both effects is investigated.

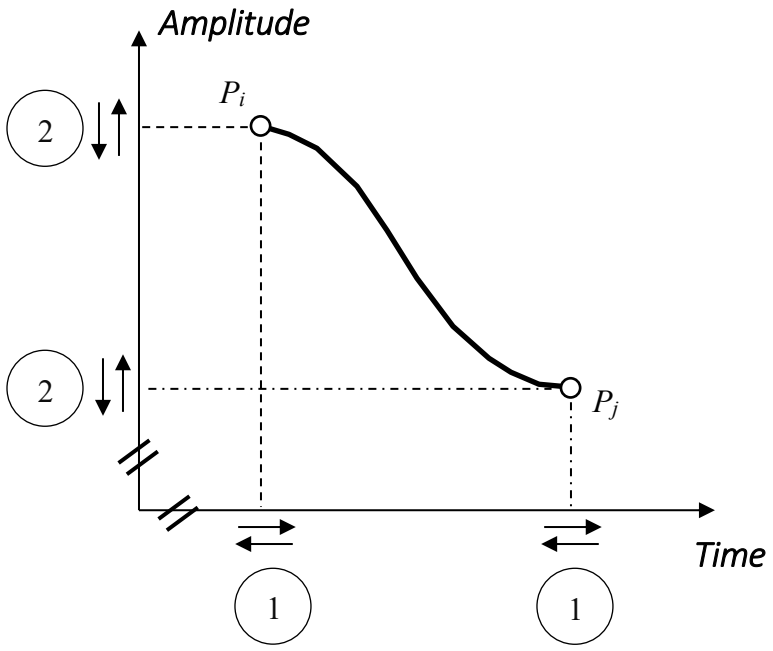


Figure 3-1: Two major effects of noise on (1) extremum point displacement, (2) fitting function fluctuation.

3.2 Signal Quality Assessment

The neural spikes studied in this work are of specific waveshapes and contaminated with background noise. As a result of the spike compression technique introduced in this work, not only the spike waveshape is altered by curve fitting to some extent, most of the spikes are also subject to considerable noise reduction. Therefore, before continuing with the details of the noise effect on the proposed spike compression approach, it is of crucial importance to introduce new measures for the assessment of the quality of the signals (neural spikes) before and after they are processed.

3.2.1 Relative Added Dissimilitude

Introduction of a new measure for reconstruction performance assessment requires a brief neuroscientific background, which is provided as follows:

In neuroscience, it is believed that the action potentials generated by a given neuron are almost the same in waveshape at the origin. When they are recorded extracellularly, however, the background noise causes slight amplitude differences, which turn those identical spikes into a 'class' of spikes. The 'within-class variability' caused by the background noise is usually ignored and all the spikes in the class under study are considered of the same origin when it comes to interpreting the information conveyed by those spikes [3]. Suppose that the neural spike extracellularly acquired by a brain-implantable neural recording device, $s_i(k)$, is the *i*-th firing of the neuron under study, where *k* is the sample index, ranging from 1 to *K*. This spike, as explained above, is assumed to be a *noiseless spike* at the origin, $s^{\dagger}(k)$, contaminated by the *i*-th observation of a random noise process, $n_i(k)$:

$$s_i(k) = s^{\dagger}(k) + n_i(k). \tag{3.1}$$

The noise is assumed to have a mean value of zero and a standard deviation of σ_N. The dissimilarity between the recorded noisy spike and the noiseless spike, which is referred to as their *dissimilitude* hereafter, is defined as:

$$d_i = \sqrt{\frac{1}{K}\sum_{k=1}^{K}[s_i(k) - s^{\dagger}(k)]^2} = \sigma_N. \tag{3.2}$$

As illustrated in Figure 3-2(a), as long as the noise standard deviation remains the same, the locus of all the spikes of the same origin recorded under the same conditions is a circle of radius σ_N (standard deviation of noise) with the associated noiseless spike, $s^{\dagger}$, at the center. In other words, despite their sample-to-sample, noise-induced amplitude differences, all such spikes have the same dissimilitude with the noiseless spike at the centre. Let us assume that a spike processing task (smart downsampling followed by proper curve fitting in our case) converts recorded noisy spikes, ***s***, to reconstructed spikes, ***r***. As a result, the dissimilitude of the spike under study with the associated noiseless spike at the centre changes from $d(=\sigma_N)$ to $\hat{d}$. Traditionally, the degradation of the spike waveshape caused by the compression-reconstruction process is assessed by the *reconstruction error* [22], [64].

$$e = \sqrt{\frac{1}{K}\sum_{k=1}^{K}[s(k) - \hat{s}(k)]^2}, \tag{3.3}$$

and can also be expressed in terms of the dissimilitudes of the original and reconstructed spikes with the noiseless spike (*i.e.*, d and $\hat{d}$, respectively) as:

$$e = \sqrt{|(d)^2 - (\hat{d})^2|}. \tag{3.4}$$

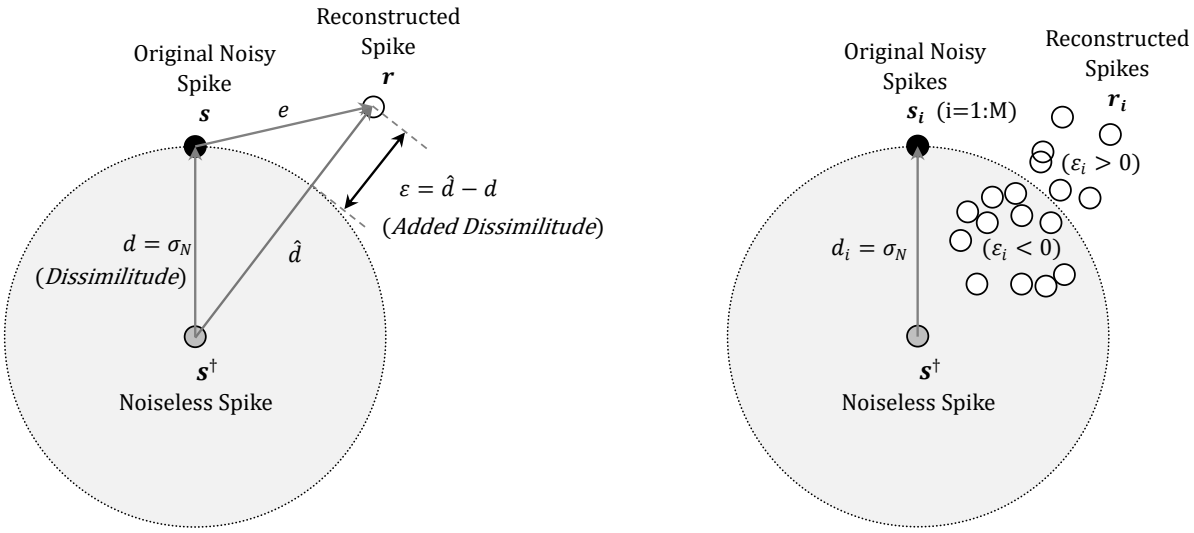


Figure 3-2: Assessment of signal quality (a) The traditional reconstruction error, e, versus the proposed 'added dissimilitude', ε. The former measures the dissimilarity of the reconstructed spike to the original noisy spike, and the latter quantifies the impact of spike processing on the dissimilarity of the spike under study with respect to the associated noiseless spike (as a reasonable reference). (b) A spike processing task, in general, might be able to reduce the noise content of a recorded spike. In this case, while the traditional reconstruction error is always positive and misleadingly reports signal quality degradation, the added dissimilitude will be negative (ε<0) indicating signal quality enhancement (*i.e.*, denoising).

The important point here is that while the reconstructed spike is away from the original noisy spike by e, what matters is still the resemblance of the reconstructed spike to the associated

noiseless spike, not to the noisy spike! Therefore, to assess the real impact of the compression-reconstruction process on the waveshape of the spike being processed, it is proposed to measure the change it causes in the dissimilitude between that spike and the noiseless spike. Referred to as the *relative added dissimilitude* (*RAD*), this is simply the relative difference between the dissimilitudes of the original and reconstructed spikes with the noiseless spike, denoted in Figure 3-2(a) as ε:

$$\varepsilon\% = [(\hat{d} - d)/d] \times 100\% \tag{3.5}$$

There is yet another reason for the superiority of the proposed 'added dissimilitude', ε, to the traditional 'reconstruction error', e (given in (3.3) and (3.4)), which is highlighted when the signal processing task to some extent denoises the spike (as it is the case in this work). According to the definition provided precedingly, denoising reduces the dissimilitude between the spike being processed and the noiseless spike, and the added dissimilitude, therefore, takes on negative values. This is while the traditional measure (e) is blind to denoising and misleadingly reports an always positive reconstruction error in such scenarios.

3.2.2 Net Added Dissimilitude

It is true that the proposed spike compression technique fits smooth curves to the salient samples of the spikes being compressed, and that provides the opportunity of reducing the unwanted amplitude fluctuations caused by the noise that is mixed with the signal upon recording. The curve fitting error together with the random nature of the added noise, however, make it difficult to claim that our spike compression always brings the recorded spikes closer to the associated noiseless spike than they were prior to compression. Therefore, the spike denoising property of the proposed idea is statistically studied for a class of M recorded noisy spikes of the same origin. As Figure 3-2(b) illustrates, let us assume that some of the reconstructed spikes fall inside the dotted circle (meaning that they have come closer to the associated noiseless spike), and some others are not. To gain an overall sense about the denoising property for the entire spike class under study, the *Net Added Dissimilitude (NAD)* is defined as:

$$NAD\% = \frac{1}{M}\sum_{m=1}^{M} \varepsilon_m \times 100\% \tag{3.6}$$

where *M* is the number of spikes in the class under study. A negative NAD% is an indication of the reconstructed spike ensemble being of smaller average distance to the noiseless spike than the

original noisy spikes are. In this case, it can be claimed that, overall, spike denoising has been accomplished.

3.3 Impact of Noise

According to the above discussions, as long as the spike under study is noiseless, selective downsampling and subsequently fitting smooth curves to the resulting salient samples works quite well. In reality, however, the non-negligible amplitude noise that is unavoidably mixed with the spike inevitably affects the idea proposed in this work. This section studies two major effects of the existence of noise: (a) extremum point displacement, and (b) the impact of noise-related salient sample amplitude fluctuations on the fitted functions.

3.3.1 Noise-Induced Extremum Point Displacement

Assuming that the locations of the start and end samples are not affected by noise, our discussion here is focused on how the introduction of amplitude noise affects the locations of extremum points. Perhaps the most critical impact of the background noise is the *displacement of extremum samples* for the spike under study. The basis for the analysis of this displacement is presented in Figure 3-3. A part of a noiseless spike containing an extremum point (point G) is shown as a solid line on the right-hand side of Figure 3-3. Let us assume that each and every

sample of this spike is subject to Gaussian amplitude noise with zero mean value ($\mu_N = 0$) and standard deviation of σ_N. As shown on the left-hand side of Figure 3-3, the amplitudes of the extremum sample G and all other samples including the neighboring samples F and H, will therefore have Gaussian distributions with the same standard deviation (*i.e.*, $\sigma_F = \sigma_G = \sigma_H = \sigma_N$) around the associated noiseless amplitudes A_F, A_G, and A_H, respectively (*i.e.*, $\mu_F = A_F$, $\mu_G = A_G$, and $\mu_H = A_H$). As illustrated in Figure 3-3, such independent noise-induced amplitude variations can slightly change the spike waveshape (from the solid curve to the dashed curve). As a result, there is a chance one of the neighboring samples F or H becomes the extremum point rather than the sample G. The probability of the displacement of the extremum point from G(T_G, A_G) to a neighboring point, say H(T_H, A_H), can be calculated by finding the probability of a sign change for the amplitude difference: $A_D = A_G - A_H$. According to [65], the difference between two random variables each being of Gaussian distribution (*e.g.*, $P_G(\mu_G,\ \sigma_G)$ and $P_H(\mu_H,\ \sigma_H)$ in our case) is of Gaussian distribution itself, $P_D(\mu_D, \sigma_D)$, where:

$$\begin{cases} \mu_D = \mu_G - \mu_H = A_G - A_H \\ \sigma_D = \sqrt{\sigma_G^2 + \sigma_H^2} = \sqrt{2}\sigma_N \end{cases} \tag{3.7}$$

Assuming that the sample G is a maximum point, as shown in Figure 3-3, the difference A_D is positive for the noise-free scenario, and the probability of a noise-induced extremum point displacement is equal to the area under the PDF curve for $A_D < 0$. Shown in Figure 3-4, as the

hashed area, this is the probability of the sample 'H' becoming the extremum point (rather than G), and called the *probability of extremum point displacement* (P_{XD}), hereafter:

$$P_{XD} = \int_{-\infty}^{0} D(\mu_D, \sigma_D)\,\mathrm{d}x = \frac{1}{2}[1 - erf(\frac{A_D}{2\sigma_N})] \tag{3.8}$$

In (3.8), erf(.) is the Gaussian distribution error function, and the probability of extremum point displacement is noticeable (*i.e.*, $P_{XD} > 0.017$) if A_D is smaller than $3\sigma_N$.

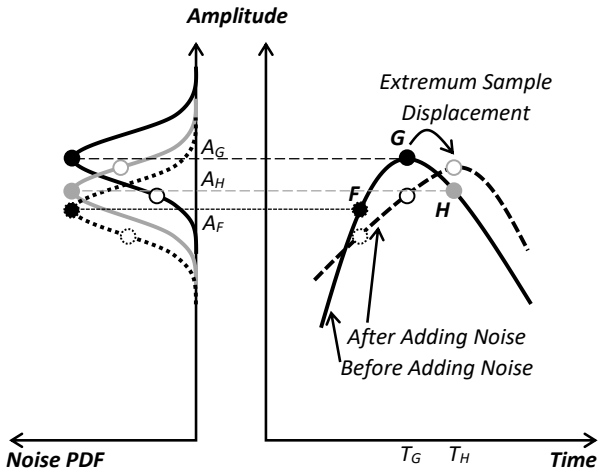


Figure 3-3: Displacement of an extremum sample as the consequence of adding noise to the spike amplitude. The extremum sample and the neighboring samples are all subject to amplitude noise. As a result, the location of the extremum sample might accordingly change.

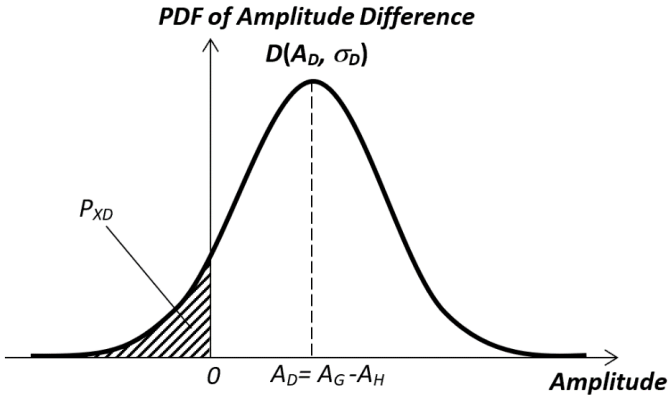


Figure 3-4: The probability of extremum points displacement (P_{XD}) can be calculated using the PDF of the amplitude difference between the extremum sample and the neighboring samples.

The chance of extremum point relocation is, therefore, contingent on both the extent of noise contamination superposed on the spike and the slope of the spike around the extremum point under study. Figure 3-5 illustrates the PDFs of displacements for two extremum points in a typical spike. So long as the noise amplitude around an extremum point is greater than $3\sigma_N$, the displacement of that point is probable with a Gaussian distribution. The higher (the absolute value of) the spike slope is around a given extremum point, the narrower the bell-shape displacement PDF will be, meaning that significant displacement of that point will be less probable.

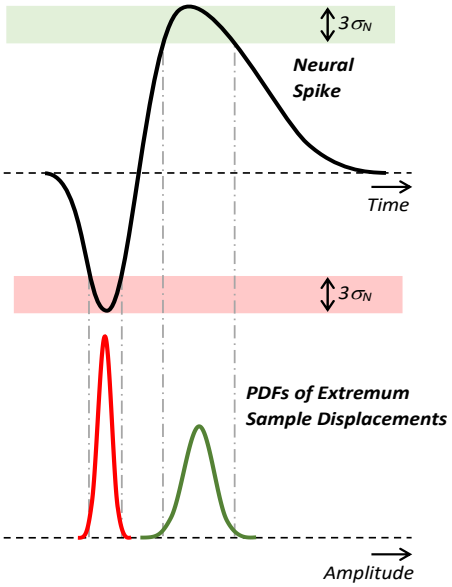
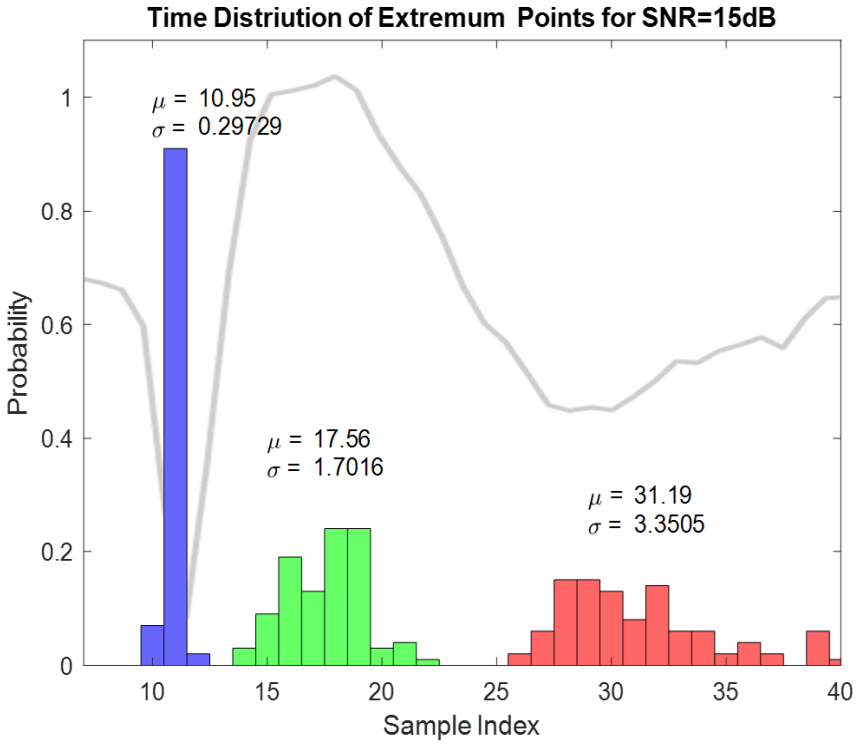


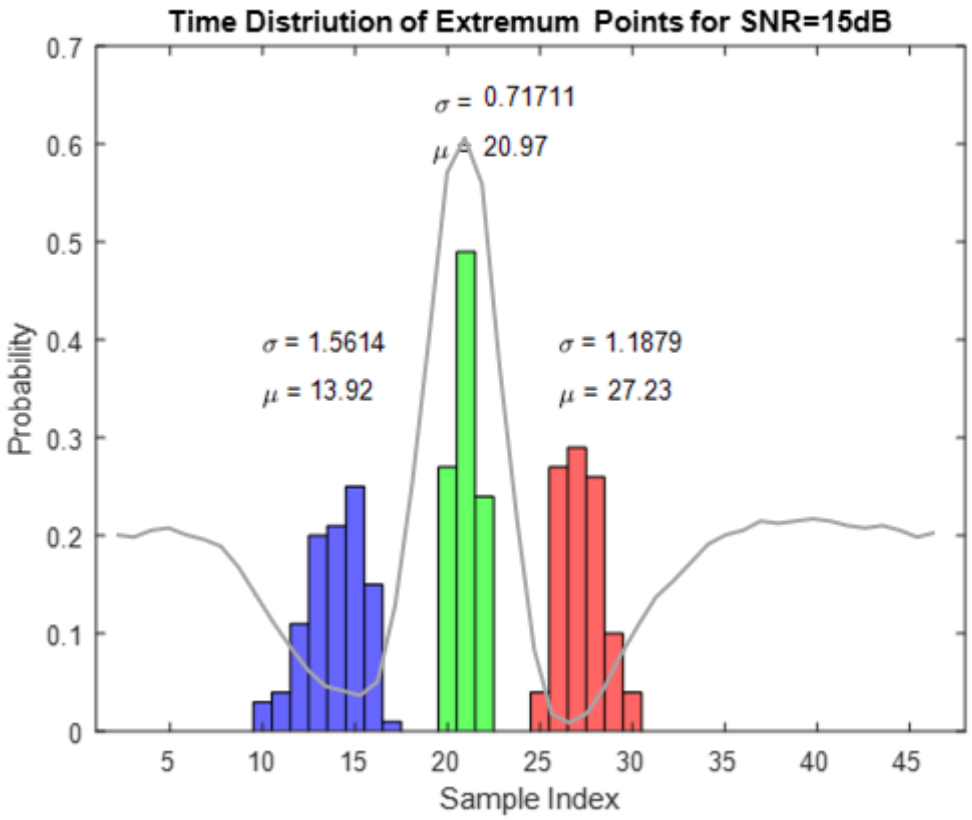
Figure 3-5: The impact of the spike slopes around extremum samples on the PDFs of extremum sample displacement in a typical spike.

Figure 3-6 shows the displacement histograms (shown in blue, green, and red separately) for the three extremum points of 100 occurrences of two actual action potentials with a *signal-to-noise ratio* (*SNR*) of 15 dB. It can be shown that such histograms would be of Gaussian form should the spike slopes on both sides of the extremum point were constant and equal in absolute value. In reality, as the histograms show, the profile of each one of the displacement distributions is *skewed* as a result of the non-constant and unequal spike slopes around the associated extremum points. Moreover, there is an inverse correspondence between the widths of the histograms and the spike slope around the associated extremum point.



(a)

Figure 3-6: Noise effect on time stamps of P_2, P_3, and P_4 for the spike waveshape shown in gray for SNR=15dB.



(b)

Figure 3-6 (continued)

3.3.2 Impact of Noise on Fitted Functions

Aside from the displacement of extremum points in time, noise-induced fluctuations in the amplitude of salient samples also cause changes in the fitted functions. To study this effect, as illustrated in Figure 3-7, let us assume that the amplitudes of the salient samples at both ends of a

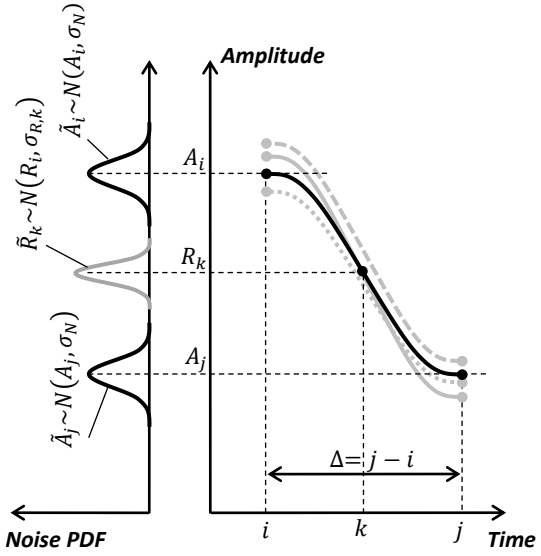


Figure 3-7: The impact of amplitude noise at both ends of a segment on the fitting functions used to reconstruct the segment waveshape

given spike segment, A_i and A_j, are contaminated with uncorrelated random Gaussian noise with a mean value of zero and standard deviation of σ_N ($N(0, \sigma_N)$). Therefore, the amplitudes of the noisy

salient samples are expressed as $\tilde{A}_i \sim N(A_i, \sigma_N)$ and $\tilde{A}_j \sim N(A_j, \sigma_N)$. Since the polynomial coefficients are functions of a linear combination of salient sample amplitudes (*e.g.*, ref. (2.5)), this noise causes Gaussian noise in the coefficients of the associated fitting function. It can be shown that the coefficients of the two-term fitting function in the presence of noise are expressed as:

$$\tilde{C}_3 \sim N(\frac{2}{\Delta^3} A_{ji}, \frac{2\sqrt{2}\sigma_N}{\Delta^3}) \tag{3.9}$$

and

$$\tilde{C}_2 \sim N(\frac{-3}{\Delta^2} A_{ji}, \frac{3\sqrt{2}\sigma_N}{\Delta^2}) \tag{3.10}$$

where $A_{ji} = A_j - A_i$ and $\Delta = j - i$. In this case, it can be shown that the amplitude of a given reconstructed sample (say sample #*k*) calculated using the resulting fitting function is subject to Gaussian noise, formulated as:

$$\tilde{R}_k \sim N(R_k, \sigma_{R,k}). \tag{3.11}$$

In (3.11), *k* is the sample index ranging from *i* to *j* (which are the indices for the start and end samples of the segment under study, respectively), R_k is the reconstructed amplitude for sample #*k* prior to adding noise, expressed as:

$$R_k = \left(\frac{2k^3}{\Delta^3} - \frac{3k^2}{\Delta^2}\right) A_{ji}, \tag{3.12}$$

and $\sigma_{R,k}$ is the standard deviation of amplitude variations for the reconstructed sample #k, formulated as:

$$\sigma_{R,k} = \sigma_N \sqrt{\left(\frac{2k^3}{\Delta^3} - \frac{3k^2}{\Delta^2} + 1\right)^2 + \left(-\frac{2k^3}{\Delta^3} + \frac{3k^2}{\Delta^2}\right)^2} \tag{3.13}$$

To analyze the effect of noise on the reconstruction error, the *Normalized Reconstruction Error Power (NREP)* is defined as:

$$NREP = NRMSE^2 = \frac{\frac{1}{\Delta}\sum_{k=i}^{j}(\tilde{R}_k - \tilde{s}_k)^2}{A_{P-P}^2} \tag{3.14}$$

Here, $\tilde{s}_k$ is the noisy amplitude of sample #k:

$$\tilde{s}_k \sim N(s_k, \sigma_N) \tag{3.15}$$

in which s_k denotes the amplitude for sample #k prior to adding noise and σ_N is the standard deviation of added noise. Hence, the reconstruction error, X_k, is formulated as the difference between the noise-induced reconstructed function ($\tilde{R}_k$) and the noisy spike ($\tilde{s}_k$) and is of Gaussian distribution:

$$X_k = \left(\tilde{R}_k - \tilde{s}_k\right) \sim N\left(E_{X,k}, \sigma_{X,k}\right) \tag{3.16}$$

in which

$$\begin{cases} E_{X,k} = R_k - s_k \\ \sigma_{X,k} = \sqrt{{\sigma_{R,k}}^2 + \sigma_N^2} \end{cases} \tag{3.17}$$

Equation (3.17) can also be written as a Gaussian distribution in the standard form (*i.e.*, with unity standard deviation):

$$X_k = \sqrt{{\sigma_{R,k}}^2 + \sigma_N^2}\ \ (N(\frac{E_k}{\sqrt{{\sigma_{R,k}}^2 + \sigma_N^2}}, 1)) \tag{3.18}$$

Replacing the reconstruction error ($X_k = \left(\tilde{R}_k - \tilde{s}_k\right)$) by the expression provided in (3.18), the NREP introduced by (3.16) and (3.17)can be expressed as:

$$NREP = \frac{1}{\Delta A_{P-P}^2} \sum_{k=i}^{j} X_k^2 = \frac{{\sigma_{R,k}}^2 + \sigma_N^2}{\Delta A_{P-P}^2} \sum_{k=i}^{j} (N(\frac{E_k}{\sqrt{{\sigma_{R,k}}^2 + \sigma_N^2}}, 1))^2 \tag{3.19}$$

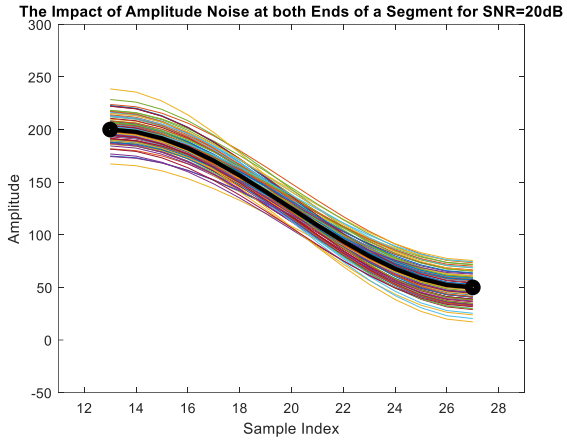
which, according to [65] and [66], is of a *noncentral Chi-squared distribution.* Noncentral chi-squared distribution has two parameters: the number of X_k terms in the summation in (3.19) called the *degrees of freedom* for the distribution (*i.e.*, $n = j - i + 1$), and λ referred to as the *non-centrality parameter* of the distribution written as:

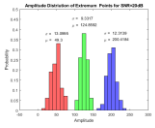
$$\lambda = \sum_{k=i}^{j} \left(\frac{E_k}{\sqrt{{\sigma_{R,k}}^2 + \sigma_N^2}}\right)^2 \tag{3.20}$$

(a) It can be shown that the total mean value and variance of the $NREP$ can be expressed as $n + \lambda$ and $2(n + 2\lambda)$, respectively.

(b)

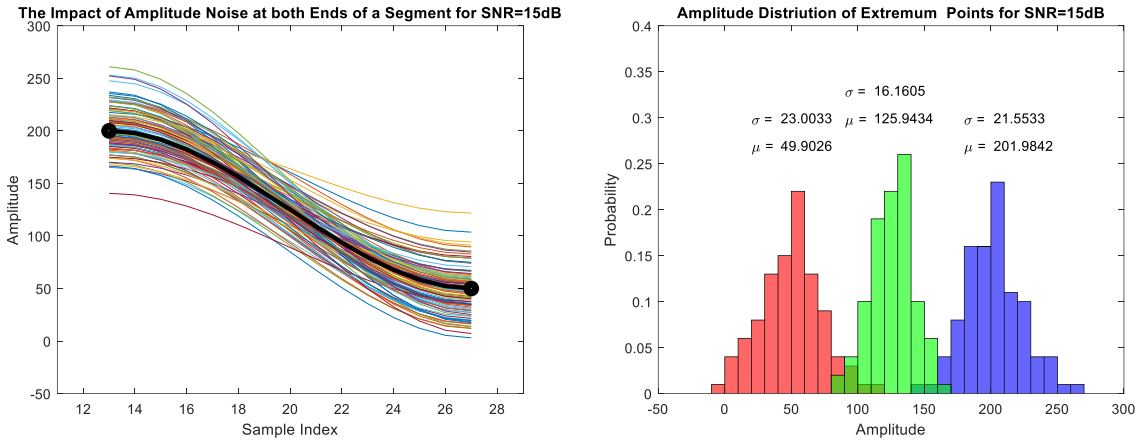
Figure 3-8, Figure 3-9 and Figure 3-10 illustrate how the random amplitude noise added at both ends of a spike segment impacts the fitted function. Assuming that the (absolute value of the) signal slope in the mid-section of a segment is greater than those at both ends of the segment, which is usually the case, standard deviation of the noise-induced fitting error in the mid-section of the segment will be smaller than that of the added noise. This is easily verifiable (a) analytically using (3.13), (b) visually by looking at each one of the time-domain segment plots, and (c) experimentally by comparing the standard deviations shown on the histograms. As a result, it can be concluded that ***statistically speaking, the overall noise-induced fitting error for the samples of the spikes reconstructed using the curve-fitting approach proposed in this research will be smaller than (for non-extremum samples) or equal to (for extremum samples) the amplitude noise contamination of the original spike.***





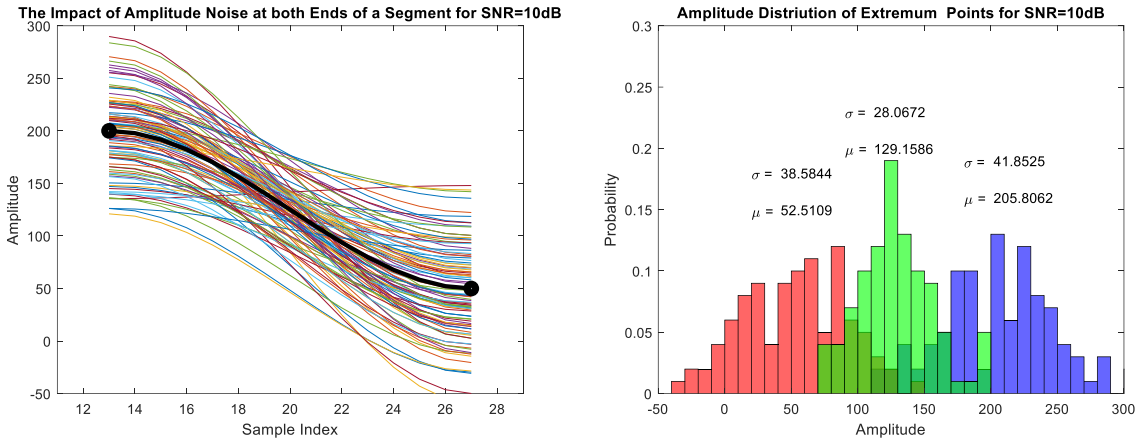
(b) (b)

Figure 3-8: The impact of amplitude noise at both ends of a segment (a) on the fitting functions used to reconstruct the segment waveshape (b) on amplitude of the fitting function (e.g., at the point with maximum slope, this point here is the middle point of the reconstructed segment), shown in the green distribution, for SNR=20dB.



(a) (b)

Figure 3-9: The impact of amplitude noise at both ends of a segment (a) on the fitting functions used to reconstruct the segment waveshape (b) on amplitude of the fitting function (e.g., at the point with maximum slope, this point here is the middle point of the reconstructed segment), shown in the green distribution, for SNR=15dB.



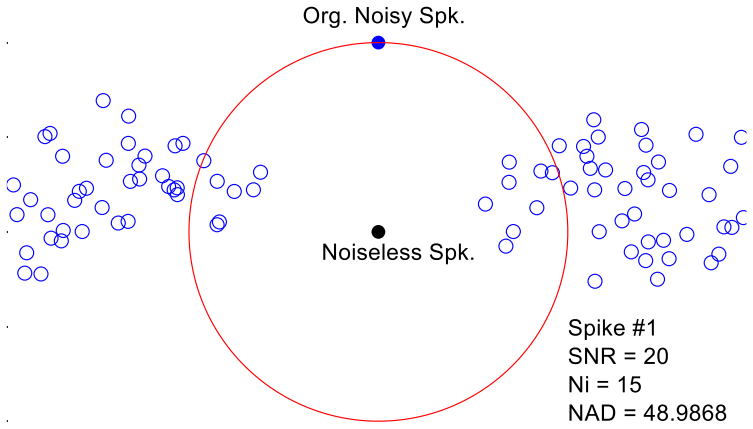
(a) (b)

Figure 3-10: The impact of amplitude noise at both ends of a segment (a) on the fitting functions used to reconstruct the segment waveshape (b) on amplitude of the fitting function (e.g., at the point with maximum slope, this point here is the middle point of the reconstructed segment), shown in the green distribution, for SNR=10dB.

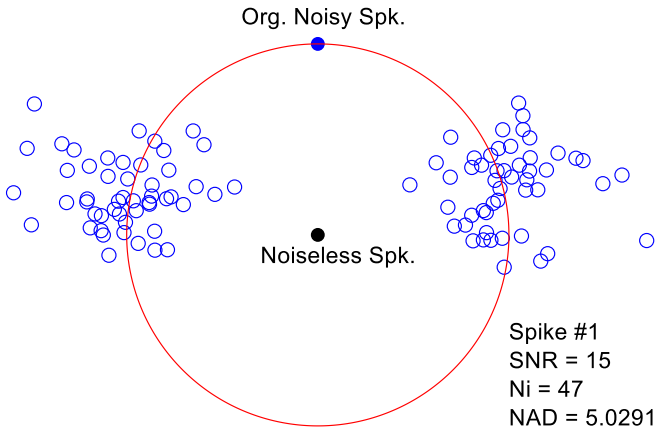
3.4 Spike Denoising

To study the denoising property of the proposed spike compression method, for each one of the waveshapes shown in Figure 2-7, classes of 100 prerecorded extracellular neural spikes with different *SNRs* are formed. For each spike waveshape, the SNR ranges from 5dB to 25dB. Figure 3-11 and Figure 3-14 show the 'dissimilitude diagram' reflecting the results of this study for Spike#1 and 4 for SNR = 20, 15, 10, and 5dB. All radiuses are normalized to the noise standard deviation, with all the 100 noisy spikes aligned on the same spot shown as a small blue filled circle, and the reconstructed spikes shown as small white filled circles. Interestingly, for spike#4 at

SNR=15dB, in 89% of the reconstructed spikes are located inside the red circle for specific SNRs, indicating negative added dissimilitudes. In other words, in 89% of cases, the compression-reconstruction approach exhibits the side benefit of spike denoising as well. The NAD for this ensemble of spikes is -22%, meaning that for this spike class overall, the proposed spike compression technique not only introduces no additional error, it even significantly reduces the overall noise contamination of the whole spike class (by 22%). The results pertaining to the other waveshapes, spike#2, 3, 5, 6, 7, and 8, are outlined in Appendix A. The findings demonstrate the presence of two different extremums. The first extremum shows the denoising property in both low and high SNR scenarios. The second extremum, however, highlights the denoising capability specially in low SNR situations, where the significance of denoising becomes more pronounced.

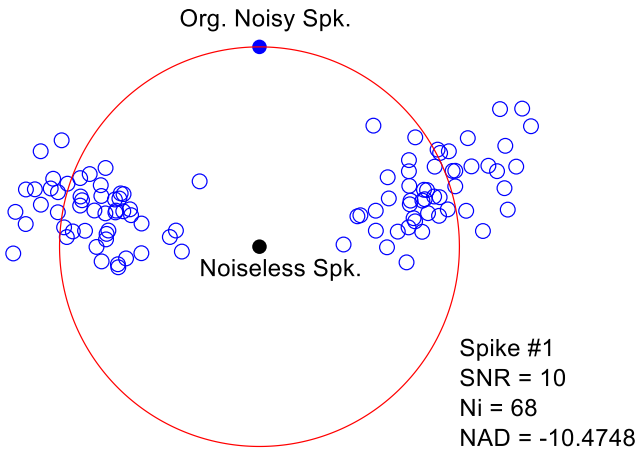


(a)



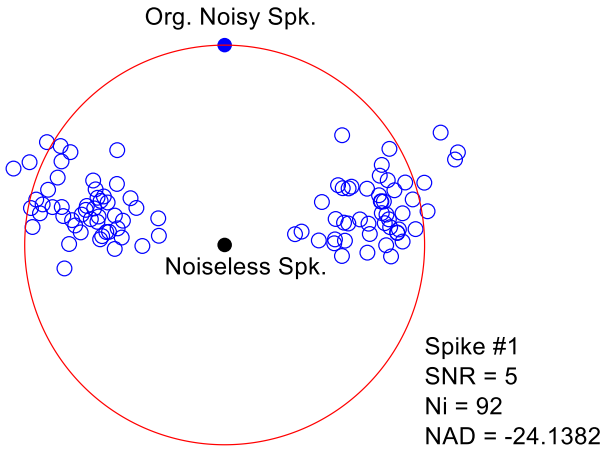
(b)

Figure 3-11: The denoising property of the proposed spike compression method. The dissimilitude diagram for a class of spikes with waveshape #1 (according to Figure 2-7); (a)-(d) for SNR=20, 15, 10, and 5dB.



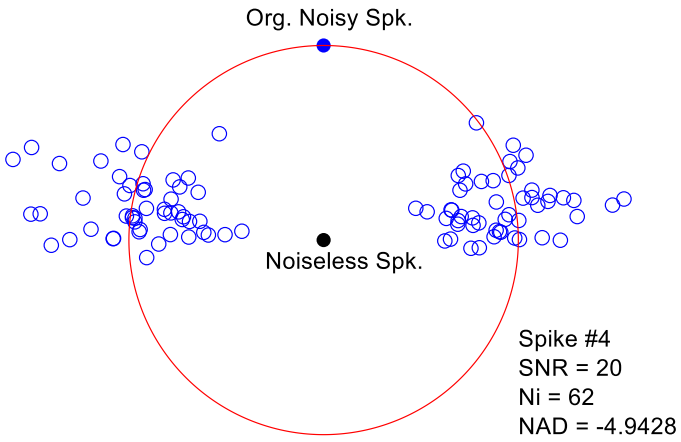
(c)

Figure 3-12 (continued)

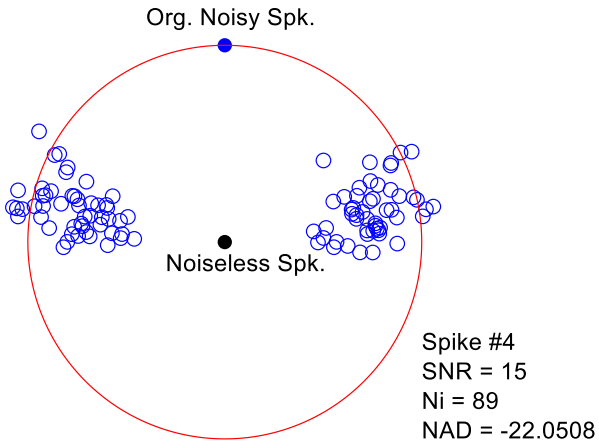


(d)

Figure 3-13 (continued)

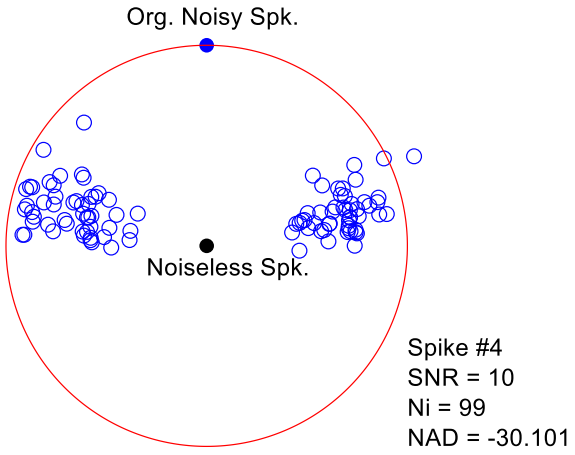


(a)



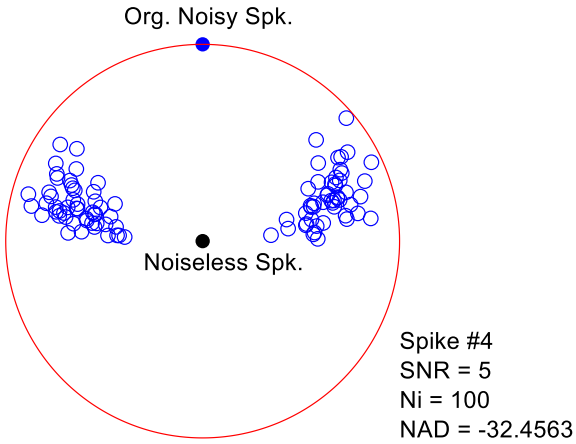
(b)

Figure 3-14: The denoising property of the proposed spike compression method. The dissimilitude diagram for a class of spikes with waveshape #1 (according to Figure 2-7); (a)-(d) for SNR=20, 15, 10, and 5dB



(c)

Figure 3-15 (continued)



(d)

Figure 3-16 (continued)

Figure 3-17 shows the NADs calculated for all the spike waveshapes in the dataset at a wide SNR range, from 5dB to 25dB. The NADs that fall below the zero level (shown using the dashed line) indicate denoising.

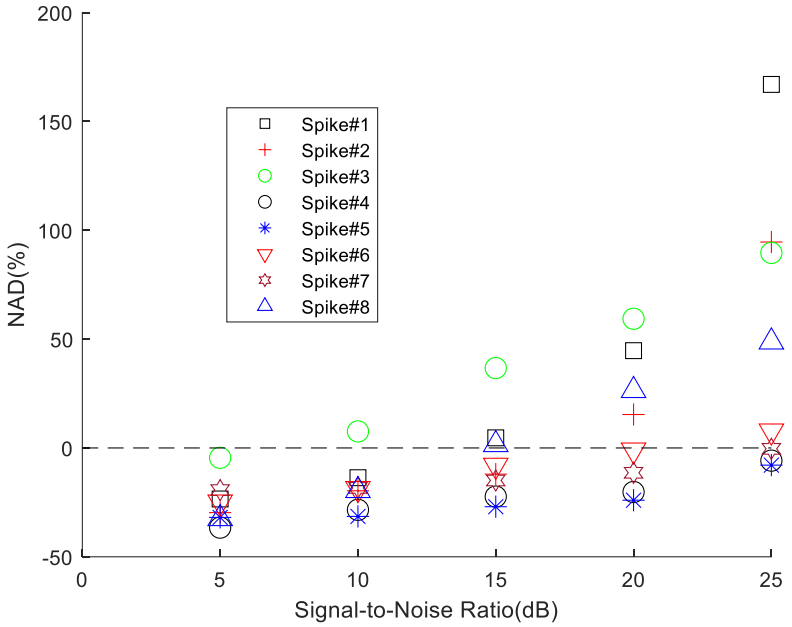


Figure 3-17: The NADs for all the 8 spike waveshapes as a function of SNR

3.5 Conclusions

This chapter discusses two primary consequences of noise presence: (a) displacement of extremum points, and (b) the impact of noise-related variations in sample amplitudes on the fitted

functions. The spike compression technique introduced in this work alters the spike waveshape to some degree through curve fitting. Moreover, a significant reduction in noise is observed in most of the spikes due to this technique. Hence, this chapter introduces new assessment measures to evaluate the signal quality (neural spikes) both before and after the processing stage. Moreover, this chapter illustrates how the compression-reconstruction approach exhibits the side benefit of spike denoising as well. Furthermore, this chapter also elaborates on the influence of noise on the displacement of extremum points and the resulting impact on the fitted functions.

Chapter 4 Hardware Design and Implementation

4.1 Introduction

To realize a multi-channel neural spike compressor based on the proposed idea, the associated hardware needs to comply with both signal processing requirements and hardware implementation restrictions. From the perspective of signal processing, the compressor circuit is expected to do the processing in the real time with acceptable precision. On the other hand, from a physical implementation standpoint, the hardware is required to be small in physical size and efficient in power consumption.

4.2 A Single-Channel Spike Compression Engine

In order to be embedded in a high-density neural recording microsystem, a single-channel spike compression engine is designed based on the proposed approach. Figure 4-1 presents a functional block diagram and timing diagram of this engine. Incoming samples of the input neural signal are grouped into consecutive non-overlapping triplets. The sum (average) of the amplitudes of the three samples in each triplet form a low-pass filtered version of the spike (represented using

triplet representative samples, TRS), which is of less noise contamination compared to the original spike. The slope of the noise-reduced signal 'TRS' is measured in the discrete time and is taken as the basis to identify the extremum samples of the spike. Amplitudes and timings of the extremum samples are then stored in a local memory referred to as the 'Salient Sample Attribute Storage'.

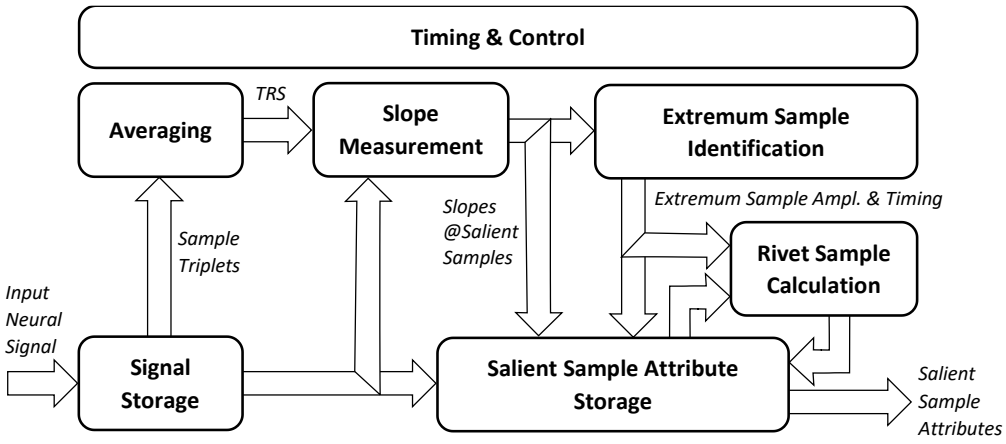


Figure 4-1: Functional block diagram of the single-channel spike compression engine implementing the proposed spike compression approach.

Figure 4-2 shows timing diagram of this engine. As mentioned before, the part of the spike that lies between every two consecutive salient sample are taken as a spike segment. If the segment spans a rather wide amplitude range (half the full-scale range in this work), slopes at both ends of the segment are calculated and stored in the Salient Sample Attribute Storage. After the completion of the spike, attributes of all the salient samples are framed in the form of a serial data packet.

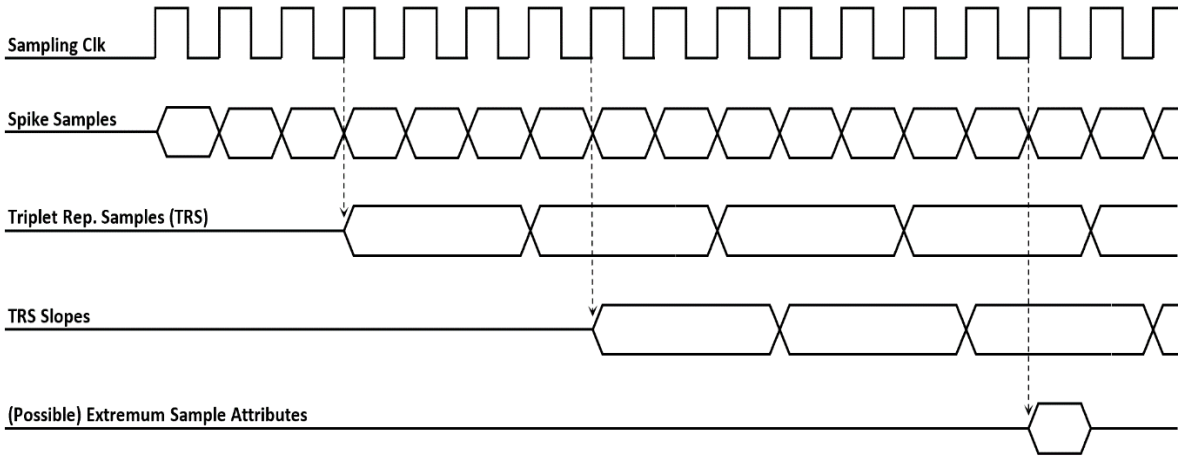


Figure 4-2: Simplified timing diagram of the single-channel spike compression engine implementing the proposed spike compression approach.

4.3 The 128-Channel Spike Compressor

Based on the proposed spike compression technique, a 128-channel spike compressor is designed. Figure 4-3 shows a simplified block diagram for this spike compressor, which comprises N single-channel spike compressors of the type shown in Figure 4-1. Prior to this circuit, neural signals on all the 128 channels are preconditioned in the analog domain, converted to digital, and then undergo a spike detection process. What is delivered to the circuit is a time-multiplexed stream of extracted neural spikes in the real time. The channels having spikes on them are referred to as ‘active channels. An active channel router receives the isolated spikes, and directs them to

one of the single-channel spike compressors. In the end, one shared data framing block collects and frames the information prepared by the spike compressor engines. In the case of having concurrent spikes on different channels, they are treated on a first come, first packed basis.

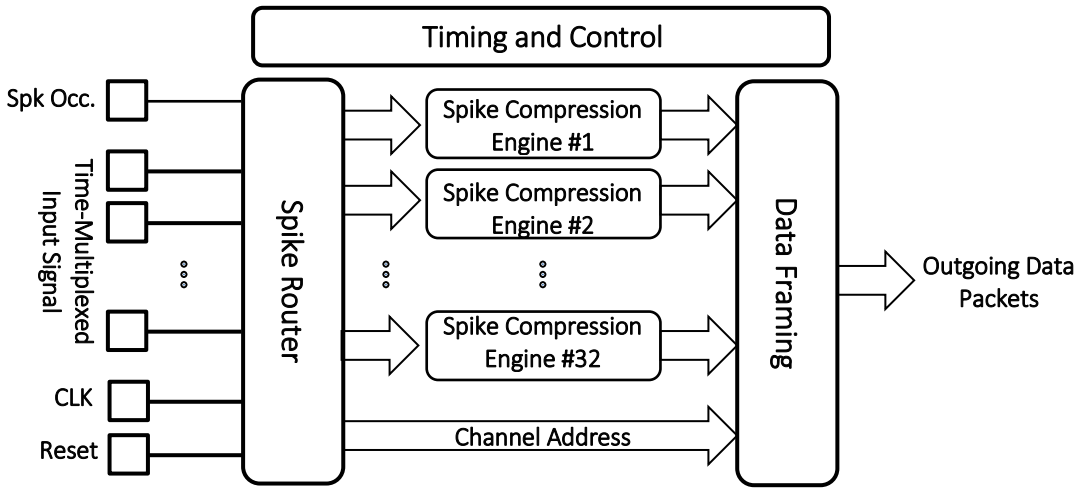


Figure 4-3: Simplified block diagram of the 128-channel spike compressor.

Figure 4-4 demonstrates the simplified block diagram of the spike router shown in Figure 4-3. Assuming that all channels are active, the timing diagram for the operation of the spike router in the 128-channel spike compressor is shown in Figure 4-5. This is, of course, a non-realistic case, for which the envisioned spike compression engines are insufficient. In reality, at a given instant of time, some channels are active and some others are not. In this case, the *spike occurrence* (*Spk*

Occ.) signal determines which channel is active and the corresponding sample is applied to the input of one of spike compression engines as shown in Figure 4-6.

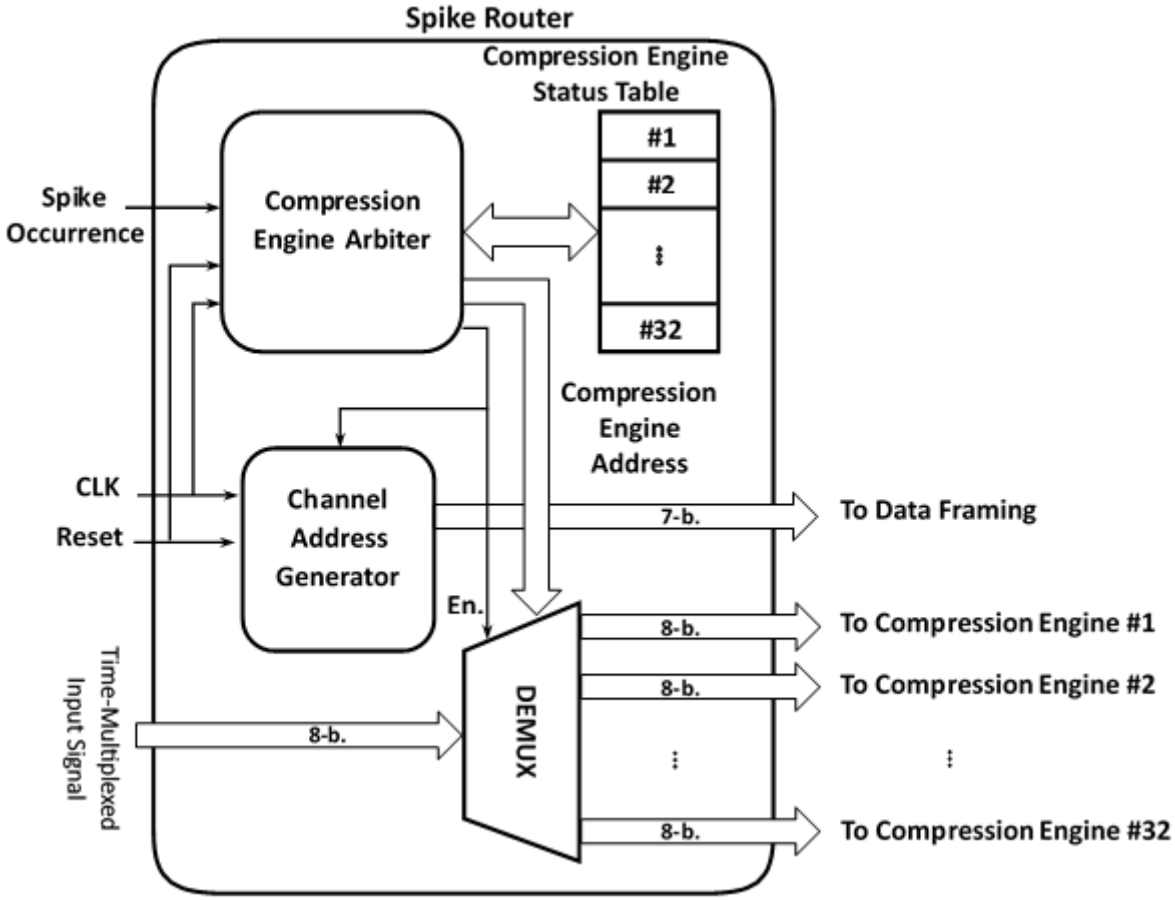


Figure 4-4: Simplified block diagram of spike router.

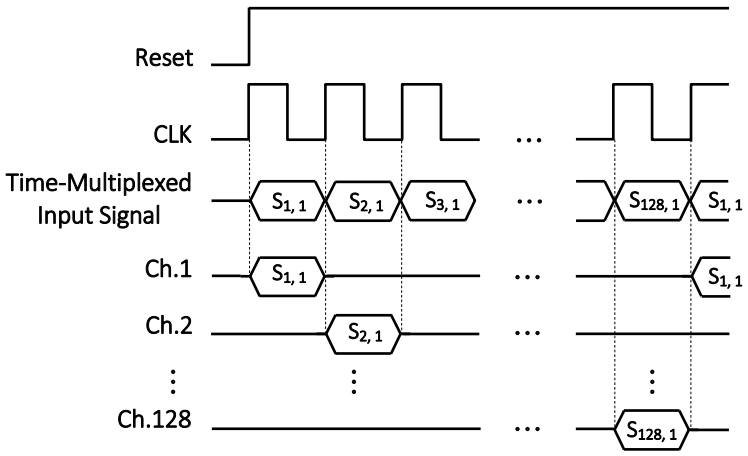


Figure 4-5: Timing diagram of spike router for the case where all channels are always active.

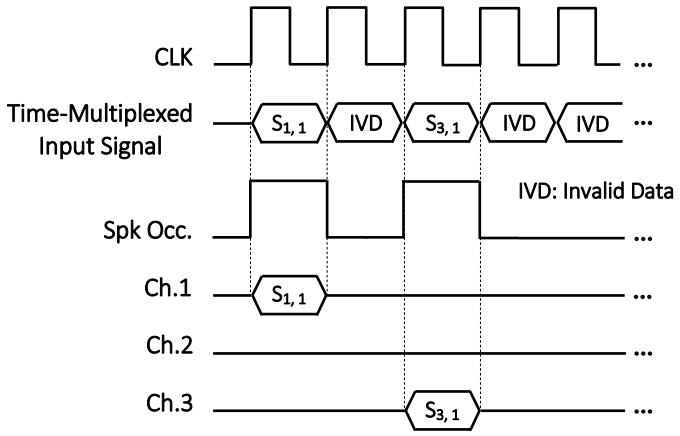


Figure 4-6: Timing diagram of spike router in realistic situations where some of the channels are active. The *spike occurrence* (*Spk Occ*.) signal shows whether the received data is a spike sample or the received data is invalid (IVD).

4.3.1 Number of Single-Channel Engines

It is believed in neuroscience that spontaneous neuronal activities occur with the Poisson distribution [30]. This is taken as the computation basis for the calculation of the number of spike compression engines (N) in the 128-channel spike compressor. The probability of having a given number of events occurred in a fixed interval of time with the Poisson distribution is formulated as:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{4.1}$$

where λ is the average number of events, and x is the number of events. The average number of events (occurrence of neural spikes in our case), λ, is calculated in our case as:

$$\lambda = FR \times T_{SPK} \times N_{ch} \tag{4.2}$$

where FR is the spike firing rate, T_{SPK} is the average time course of a typical spike, and N_{ch} is the total number of channels. For a neural signal with average spike firing rate of 100 spike/s and typical spike duration of 2ms, the average number of events on N_{ch} =128 channels equals 25.6 spikes. According to (4.1), for λ=25.6, the probability of having more than 32 concurrent active channels (X=32) is as low as 12.36% as shown in Figure 4-7. Therefore, 32 salient sample detectors are envisioned in the 128-channel processor being designed.

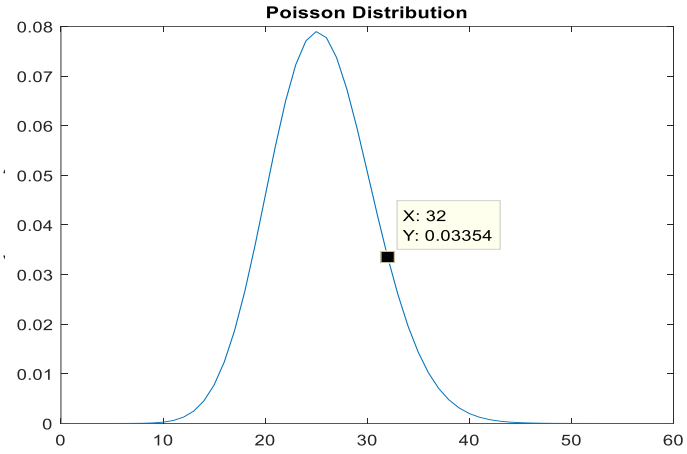


Figure 4-7: Poisson distribution for λ=25.6

4.3.2 Timing Considerations

Real-time operation of the proposed spike compressor is of crucial importance in applications such as prosthetic devices and brain-machine interfaces [67], [68]. The fact that the single-channel salient sample detector functions in the real time when operated at a clock rate equal to the basic front-end sampling rate (i.e., 25 kSample/Sec.) is translated into extremely low dynamic power consumption, especially when the system is designed for high channel counts. The only parts of the 128-channel spike compressor that is operated at a high clock rate are the active channel router and the data framing block. To guarantee live streaming of the compressed neuronal activities to the outside, these blocks receive a clock rate 128 times higher than the basic sampling clock rate.

4.3.3 The Rivet Sample Impact

To demonstrate the impact of the rivet point in reducing the fitting error, the graph in Figure 4-8 compares the curve fitting errors for all the 8 spike waveshapes with and without rivet points. Averaged on all the noiseless spikes in the dataset, the overall curve fitting error (calculated based on the normalized sample-to-sample error introduced in [64] reduces from 5.98% to 1.8% as a result of using rivet samples. The price of this significant improvement in the accuracy of the curve fitting is paid by sacrificing a fraction of the compression rate. At a firing rate of 8 spikes/sec., the compression rate (CR) drops from 390 down to 272. Normalized to the firing rate, as introduced in [22], the true compression rate (TCR) reduces from 3125 to 2176. Interestingly, even after this degradation in the compression rate, the TCR in this work is still high enough to outperform all other works so far reported in the literature (ref. Table 4-2).

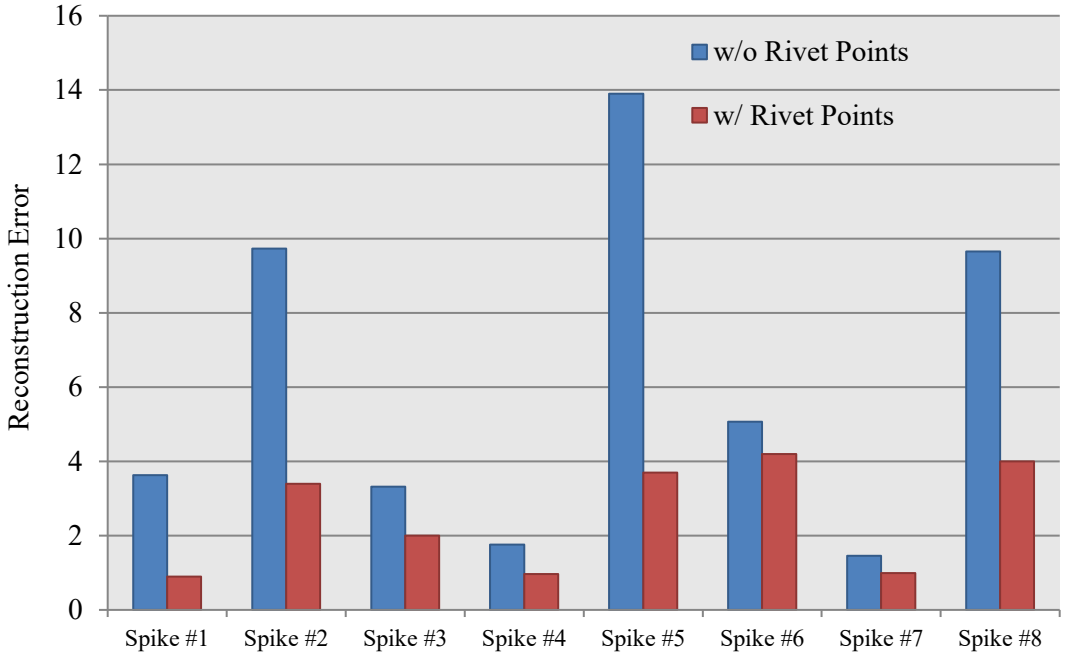


Figure 4-8: The reconstruction error without using the rivet sample and with using the rivet sample.

4.4 Outgoing Data Packets

The attributes of the salient and rivet samples acquired by the 128-channel spike compressor are framed in the form of serial data packets. Each packet starts with a start pulse and ends with a stop pulse for synchronization purposes. To allow for the live streaming of spike information, the data framing block is operated at a 128 times higher rate than the basic sampling clock. This clock

signal is taken from the active channel router, which operates at the same frequency. Moreover, to reduce power consumption, the data framing block is only activated when the data is prepared for transmission. Assuming employing on-off keying (OOK) modulation in wireless transmitter of the implant, this feature saves power significantly. Typically, with a spike course of 1 ms and firing rate of 10 spike/s, 90% power is saved. In the worst case, with the maximum firing rate ranging from 100 to 200, 80% power is saved. As shown in Figure 4-9(a), the data packet reporting a spike starts with eight 1's (*i.e.*, FF_{Hex}) as the *start pulse* for synchronization purposes follow ed by the associated channel address, *CHA*. Then follows segment information, which can be of one of the two possible formats shown in Figure 4-9(b) depending on the slopes of the segment at both ends. For each segment, the time difference between the start and end samples (T_j-T_i) as well as the amplitudes of those samples (A_j and A_i) are always reported. If the segment is modeled using a 5-term polynomial as explained before (in section 2.3.6), the associated segment field will also include the forward differences Δ_i and Δ_{j-1} and the amplitude of the rivet sample (A_{Rivet}). Upon receiving the first three fields for each segment, if the difference between start and end amplitudes is greater than half the full-scale range, the external host assumes that the segment is modeled using a 5-term polynomial, and takes the next three subfields as Δ_i, Δ_{j-1}, and A_{Rivet}, being 4, 4, and 8 bits in length, respectively. For error control, parity bits (P_1–P_3) are added to the salient sample fields as shown in Figure 4-9(b). P_1 and P_2 are odd parity bits associated with A_i and A_j,

respectively, while P_3 is an additional odd parity bit that corresponds to the remaining part of the packet.

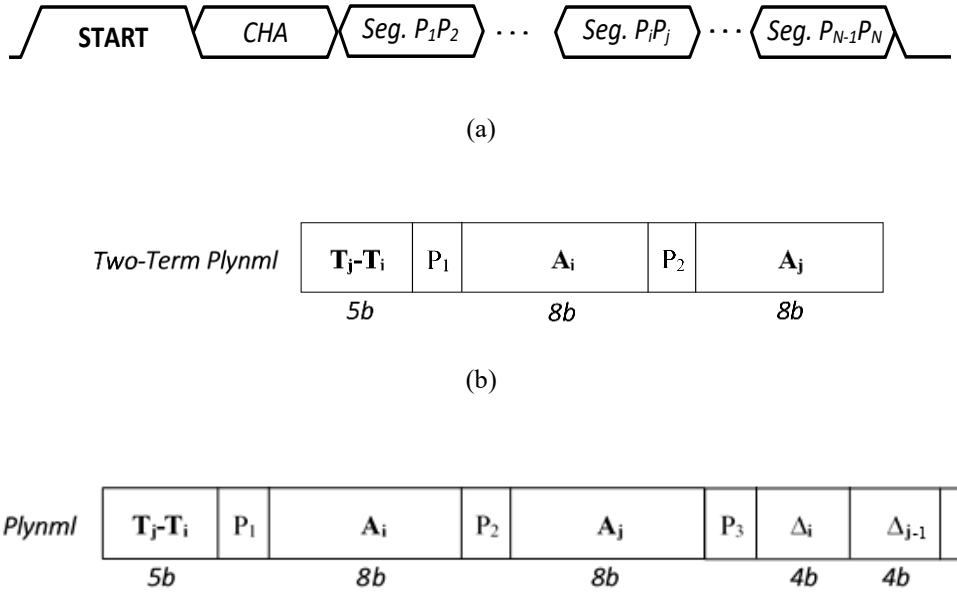


Figure 4-9: Outgoing data packet (a) General format, (b) Details of salient sample and rivet sample fields

4.5 Power-/Area-Efficient Hardware Design

Aside from signal- and system-level considerations, the efficient design of the hardware realize the proposed compression technique is of crucial importance in achieving implant-appropriate performance. The efforts put in this direction can be summarized as follows:

I. In synchronous digital circuits, a clock signal is used to synchronize the operation of different components. However, not all components need to be active and consuming power at all times. In many designs, there are portions of the circuit that are only active under certain conditions or during specific phases of operation. For instance, in order to reduce dynamic power consumption of the compressor the clock gating technique is to selectively disable (gate) the clock signal to the data framing block and the spike compression engines when they are not active, thereby reducing power consumption.

II. At the architecture level, multiple 'Salient Sample Detector' blocks are available in parallel in order to process incoming spikes on active channels concurrently. This allows for the real-time operation of the multi-channel spike compressor with no enhancement in the clock rate, hence avoiding unnecessary penalty in dynamic power consumption. Moreover, the hardware used for 'Averaging' and 'Slope Measurement' in the block diagram of Fig. 8(a) are shared among all the 32 Salient Sample Detectors. Given that the Salient Sample Detectors are operated sequentially with a duty cycle of 1:32, this makes the designed processor efficient in silicon area (13%) with no speed or power consumption penalty.

III. At the building block level, area-/power-efficient computational building blocks are used to realize the processing tasks. As examples, multiple additions in the Low-Pass Filtering blocks (for the calculation of the amplitude of triplet average samples, TRSs)

are simultaneously performed using carry-save adders, gray-code counters are used (for purposes such as sample time stamping and Salient Sample Detector address generation) to lower the consumed power.

IV. At the signal level, the entire algorithm proposed in this work is realized with no need for multipliers, resulting in savings in power and area. To avoid extra power and area consumption, the length of the data conveying amplitudes is always truncated to 8 bits. Instead of calculating spike slopes, the differences between the amplitudes of consecutive samples are calculated and transmitted off the implant module. The temporal spacing between salient samples is transmitted to the outside rather than their absolute time stamps. This results in 9% enhancement in the compression rate.

V. At the protocol level, outgoing data packets convey the minimum possible information, using which the data required for curve fitting and spike reconstruction are retrieved. For instance, salient sample type (start sample, end sample, or spike extrema), salient sample absolute time stamps, the polynomial orders for spike segments are all retrieved on the external side.

4.6 Experimental Setup and Results

4.6.1 The Single-Channel Prototype Neural Spike Compressor

To assess the effectiveness of the idea suggested in this research and evaluate the functionality of the single-channel spike compression engine in executing the signal processing task it performs, an examination of its implementation on an FPGA evaluation board is conducted. Consequently, a behavioral model of the proposed approach is created using VHDL description via the ISE Xilinx® developer. The synthesis report generated by ISE Xilinx® outlines the digital blocks included in the single-channel engine, as detailed in Table 4-1.

Table 4-1: Digital blocks used to realize the single-channel compression engine

Digital Block	**Number**
8-bit registers	148
D-type Flip flops	132
LUTs	210

Our laboratory is equipped with DE10 FPGA board. Therefore, in order to use the VHDL description developed via ISE Xilinx® developer, an Intel® Quartus® software project is initiated, along with the creation of a VHDL module. Subsequently, pin assignments are completed, the VHDL code is compiled, and finally, the DE10 FPGA board is configured. As shown in Figure 4-10, the DE10 FPGA board is a popular development platform produced by Terasic, a leading provider of FPGA-based solutions [69]. The board is designed to facilitate the development and testing of digital designs, particularly for applications involving high-performance computing, image processing, signal processing, and embedded systems.

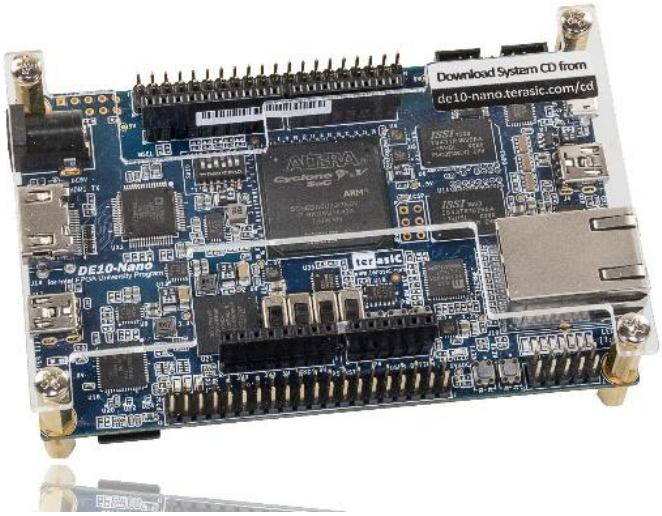


Figure 4-10: DE10-nano Development Kit used to verify the single-channel compression engine.

4.6.2 The Microfabricated Neural Spike Compressor

The 128-channel neural spike compressor designed in this work was microfabricated in TSMC 130-nm CMOS process. The physical layout of the core circuit, depicted in Figure 4-11, spans a measurement of 350 µm by 1050 µm. Additionally, Figure 4-12 shows a photograph of the chip that has been fabricated and wire-bonded.

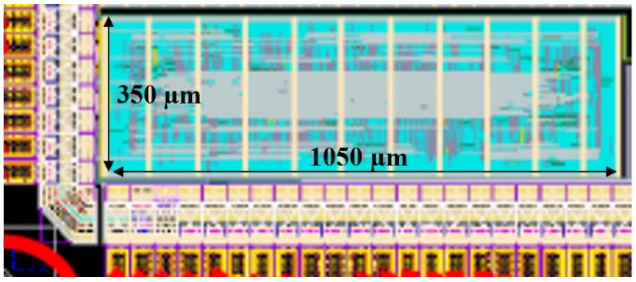


Figure 4-11: The microfabricated 128-channel spike compressor, physical layout of the core circuit

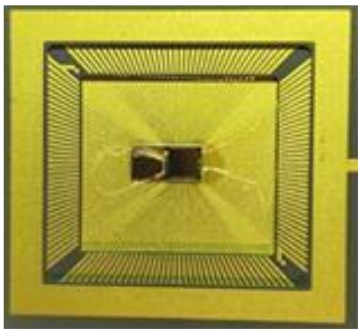


Figure 4-12: The microfabricated 128-channel spike compressor, photograph of the chip

Figure 4-13 and Figure 4-14 illustrates the experimental setup employed. To test the chip, initially, eight distinct isolated spikes, each with different SNRs, are stored as test vectors on the FPGA board. In both scenarios involving overlapping and non-overlapping occurrences, these isolated spikes are then inputted to the fabricated chip. Following this, the spike attributes are framed and reported to the external side via the chip's serial output. A data acquisition device facilitates the interface between the compressor's output and a computer. According to our measurements, operated at a supply voltage of 1V and a master clock rate of 3.2 MHz, the 128-channel processor consumes a total power of 21 μW (0.164 μW per channel). To measure power consumption, a multimeter is connected in series with the power supply and the fabricated chip, measuring the DC current (average current) of the chip. Subsequently, a scenario with 32 active channels is set up on the FPGA board, and the chip is supplied with data. The fabricated chip processes the data through 32 compression engines (fully loaded). In this scenario, the power consumption of the chip is determined by multiplying the power supply voltage and the measured DC current.

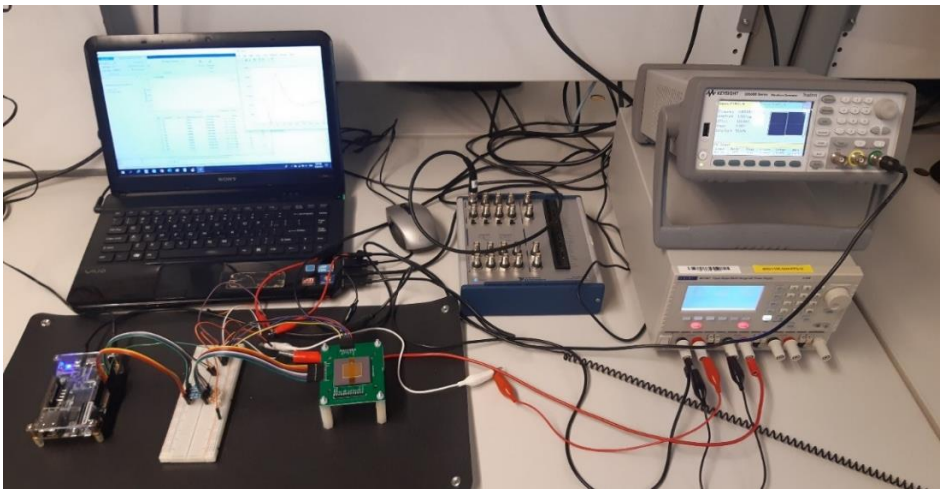


Figure 4-13: The microfabricated 128-channel spike compressor, the experimental setup

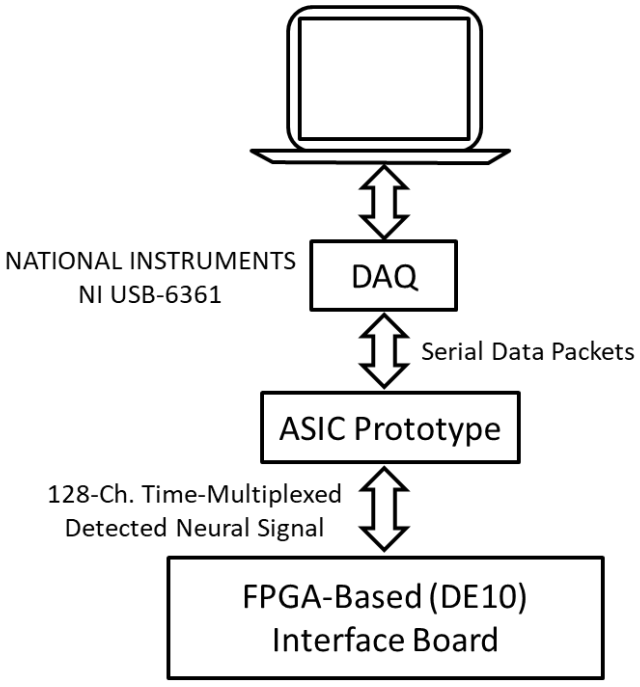
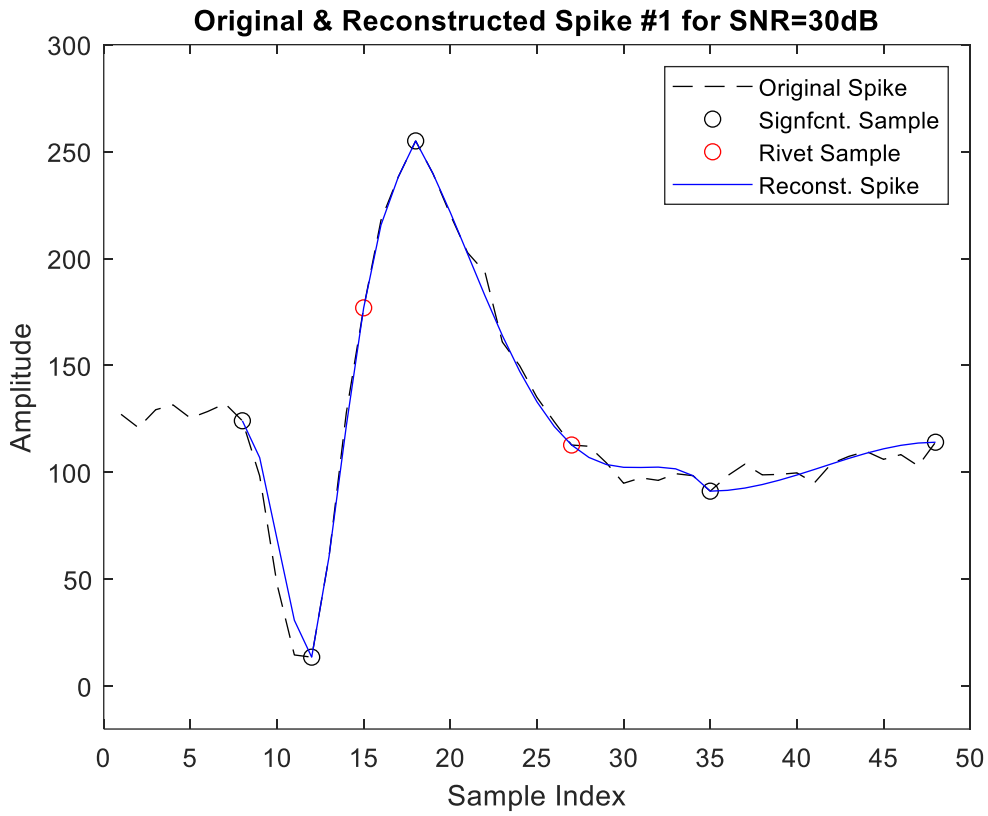


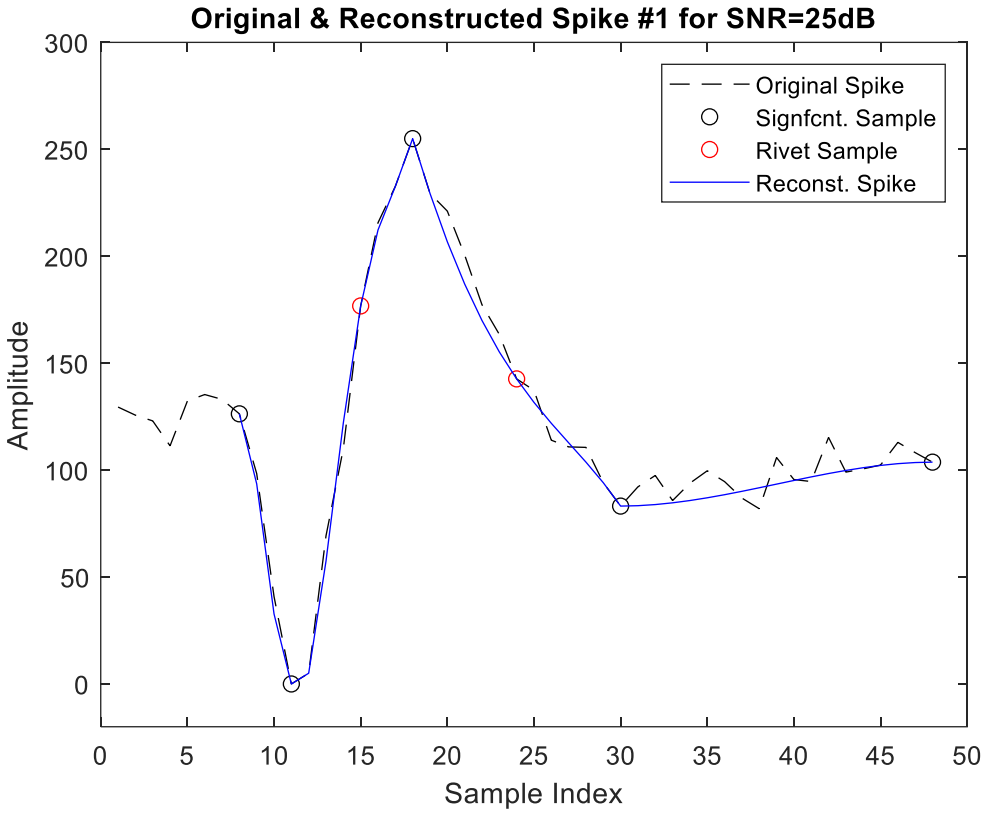
Figure 4-14: The experimental setup to test the ASIC prototype

Figure 4-15 shows the reconstructed spikes on the external side for SNR spans from 5dB to 30dB.



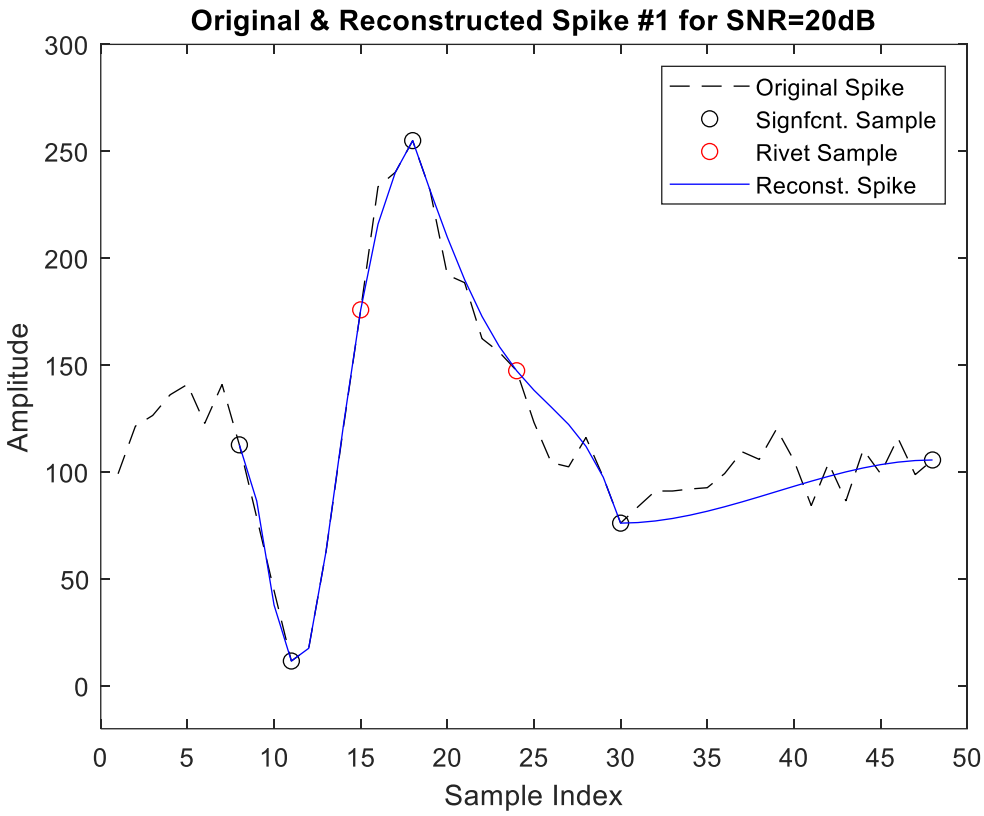
(a)

Figure 4-15: Experimental results demonstrating the proposed technique showing the original spike (dashed line), reconstructed spike (solid line), salient samples (black circles), and rivet samples (red circles); (a)-(f) present the results for spikes #1 for SNR=30dB to 5dB, respectively



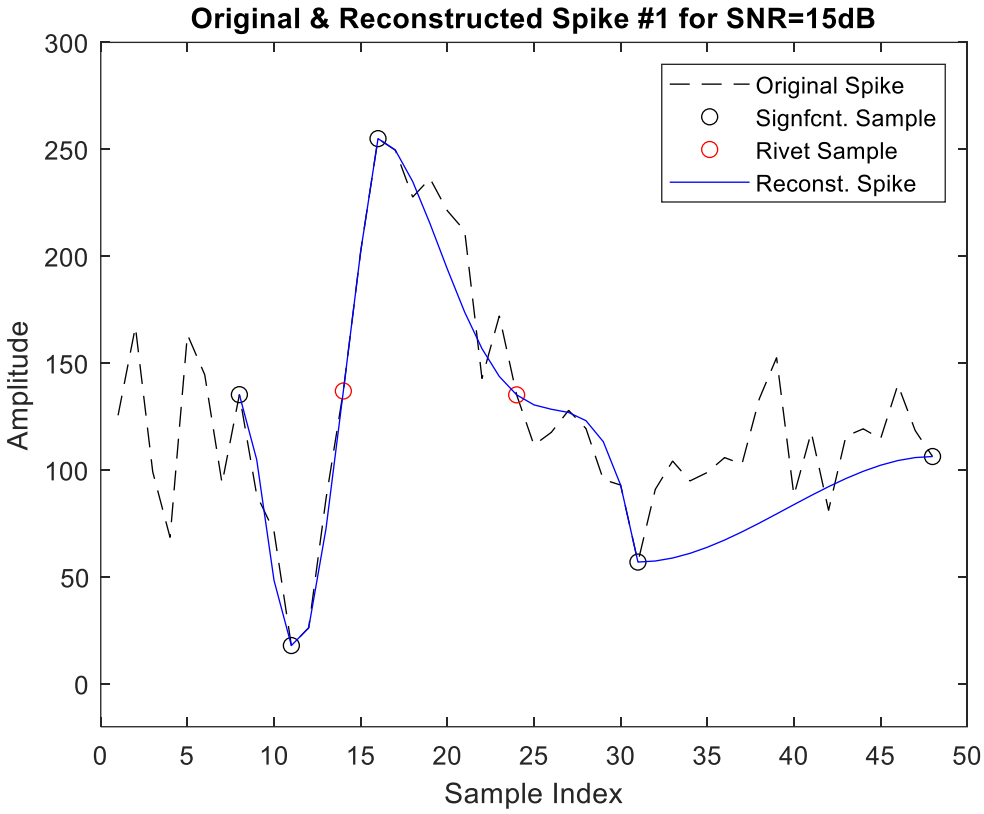
(b)

Figure 4-16 (continued)



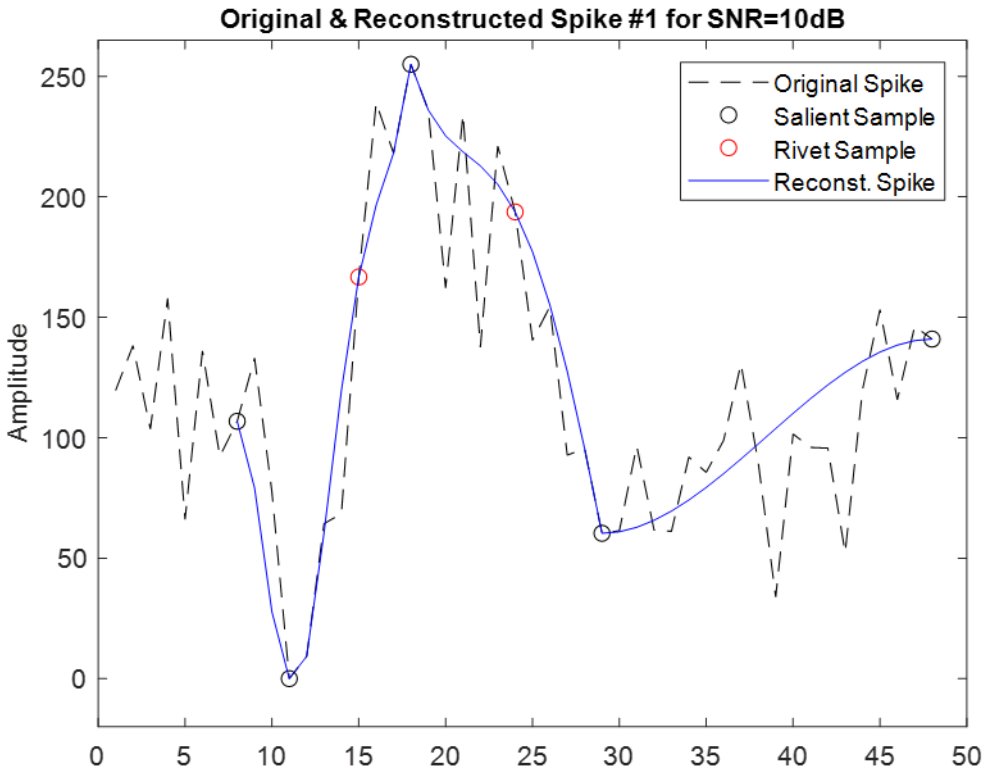
(c)

Figure 4-17 (continued)



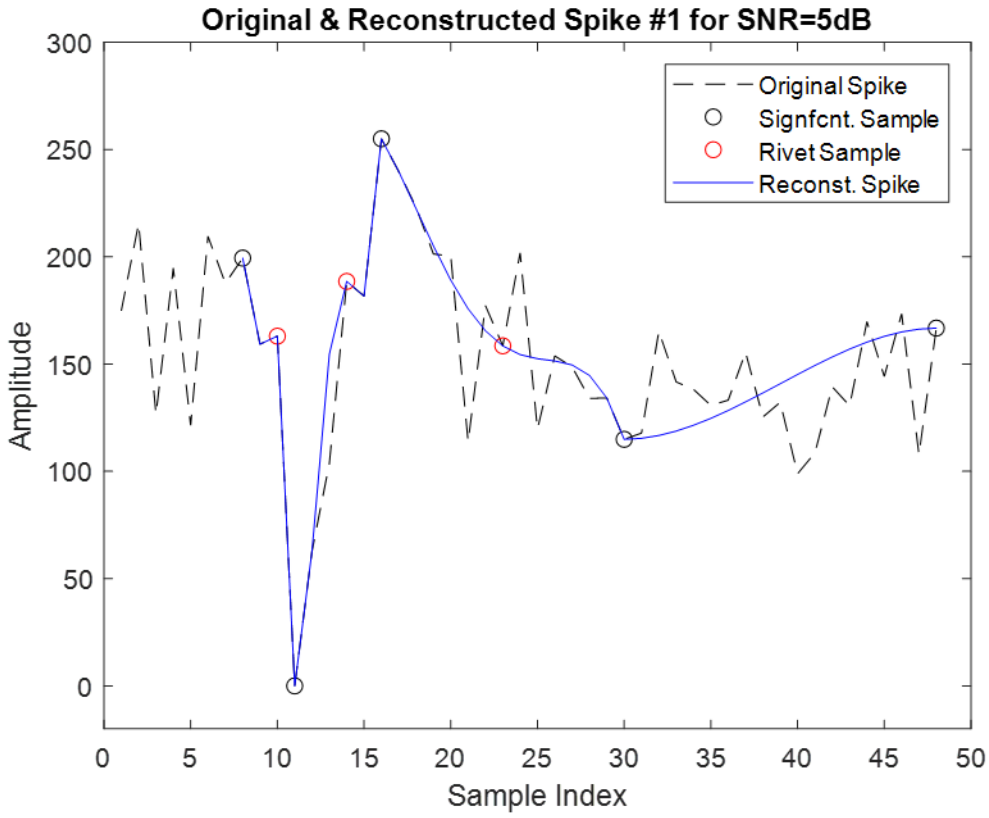
(d)

Figure 4-18 (continued)



(e)

Figure 4-19 (continued)



(f)

Figure 4-20 (continued)

The breakdowns of the overall silicon area and measured power consumed by the 128-channel spike compressor are shown in Figure 4-21. As the pie charts show, as low as only 10% of the area and 26% of the power is spent on computations in the proposed algorithm, and the rest is consumed by storage elements.

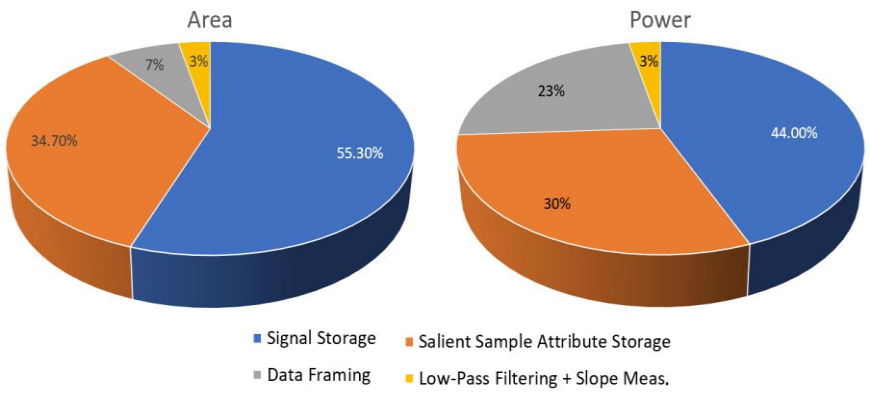


Figure 4-21: Breakdowns of silicon area and measured power for the 128-channel spike compressor

Table 4-2 summarizes the design, operational, and performance specifications of the spike compressor in this work and compares them with those of other similar works appeared in the literature. Thanks to the computationally-light technique proposed in this work, our measured per-channel power consumption is significantly lower (by 62%) than the best spike compressor reported so far, and the per-channel silicon area occupied by our processor is less than two third of that of the most compact work appeared in the literature. In addition to the superiority of the proposed spike compressor to other works in hardware implementation, its signal compression performance shows significant improvement compared to the best spike compressors ever reported. As a normalized measure for signal compression, the TCR in this work is more than twice the highest TCR appeared in the literature (2176 vs. 915).

Table 4-2: Performance summary of the proposed neural signal processer in data reduction in comparison with other works.

Math. Technique		**DPCM**	**Compressive Sensing**	**MBED**	**DWT/ HAAR**	**This Work**
Reference		[41]	[43]	[49]	[22]	-
Year		2021	2014	2020	2015	2023
Technique Type		Temporal	Spatial	Spatial	Temporal	Temporal
Technology (µm)		0.028	0.18	0.13	0.13	0.13
Supply Voltage		1	1.2	1.2	1.2	1
Sampling Rate (kS/sec.)		24	4	NA	20	25
No. of Channels		1	16	16	64	128
Firing Rate		2	NA	NA	8	8
Compression Measure	CR	163	16	48	116	**272**
	TCR	326	NA	NA	904	**2176**
Area (mm^2)		0.00079	0.128	0.048	0.206	**0.367**
Area/Ch. (mm^2)		0.00079	0.008	0.003	0.0032	**0.00287**
Power (µW)		1.05	15.2*	102.4	94.08	**21**
Power/Ch. (µW)		1.05	0.95	6.4	1.47	**0.164**

* Includes spike detection power consumption.

4.7 Design Flow and CAD Tools

In Hardware Description Language (HDL), behavioral and structural modeling are two different approaches used to describe the functionality and structure of digital circuits. Behavioral modeling focuses on specifying the functionality or behavior of a digital circuit without explicitly detailing its internal structure. It allows for a clear representation of the intended functionality,

making it easier to understand and modify the code. It is often used during the early stages of design (as we used for the single-channel prototype neural signal compressor). On the other hand, it may not provide detailed control over the physical implementation of the circuit, which can be a limitation in some cases. Structural modeling, on the other hand, involves specifying the internal structure of a digital circuit by interconnecting lower-level components. It provides a detailed representation of the hardware structure, making it suitable for detailed design and analysis. It allows for better control over the physical implementation of the circuit. With structural modeling, designers have more control over the optimization of specific hardware components, potentially leading to more efficient implementations. Therefore, a structural modeling is created using VHDL description via the ISE Xilinx® developer to implement the 128-channel neural signal compressor. It is noteworthy that, despite the availability of customized resources on FPGA such as SRAM, for an efficient ASIC implementation of the hardware, we opt to use registers as memory cells in the 128-channel neural signal compressor. Subsequently, Design Compiler® is utilized to obtain the gate-level representation of the model for automated layout generation. Innovus is employed for this purpose. Following that, Cadence® is utilized for pad framing. Ultimately, for layout verification or post-layout simulation, Ncsim® is employed.

4.8 Conclusions

This study proposes a novel technique in temporal spike compression, which is based on the segmentation of neural spikes, transmission of the key spike segment attributes off the neural recording implant module, and reconstruction of the spike waveshape by curve fitting. This and several other design techniques at signal and circuit levels help achieve the highest compression performance ever reported.

Chapter 5 Conclusions and Future Works

Neural signal compression is a field of study focused on reducing the size and complexity of neural data while preserving relevant information. The conclusions and future works in this area are constantly evolving, but I can provide you with some general insights based on the research up until my knowledge cut-off in September 2021.

5.1 Future Works

- ***Employing neural networks for more efficient curve fitting-*** Benefitting from computational capabilities of certain types of artificial neural networks (*e.g.*, autoencoders), especially curve fitting and interpolation.
- ***Standardization efforts-*** Developing standardized formats and protocols for compressed neural signals can facilitate interoperability and data sharing among different research groups. Establishing common benchmarks and evaluation metrics will also aid in comparing and assessing the performance of different compression techniques.

- ***Ethical considerations-*** With the increasing use of neural interfaces and brain-computer interfaces, ethical considerations surrounding the compression of neural signals will become important. Future works should address privacy concerns, data security, and ensure that compression techniques, especially when the recorded neural data is to be telemetered through wireless links.

5.2 Conclusions

A novel method for spike compression in brain-implantable microsystems is proposed. Unlike previous neural signal compression techniques in the literature, the proposed method does no major signal processing is performed on the implant side of the system. Instead, this work proposes a general piecewise formulation for typical spike waveshapes. This formulation will need only timing and amplitude of significant samples of the spike as well as the associated slopes in order to reconstruct the spike wave shape off the implant. Key advantage of the proposed approach compared to other existing methods is the high compression rate along with extremely simple hardware on the implant. This makes the proposed technique appropriate for incorporation in high-density neural recording microsystems.

In addition to its effectiveness in data reconstruction, noise analysis is the key requirement that define the suitability of such technique for on-implant neural signal processing. Therefore, chapter 3, first, challenges the conventional reconstruction evaluation and in advance presents two new signal quality assessments. Second, it studies two major effects of the existence of noise: (1) extremum point displacement, and (2) the impact of noise-related salient sample amplitude fluctuations on the fitted functions. At the end, the effect of both are investigated.

References

[1] K. D. Wise, A. M. Sodagar, Y. Yao, M. N. Gulari, G. E. Perlin and K. Najafi, "Microelectrodes, microelectronics, and implantable neural microsystems," *Proceedings of the IEEE,* vol. 96, no. 7, pp. 1184-1202, July 2008.

[2] M. Shaeri and A. M. Sodagar, "Data Transformation in the Processing of Neuronal Signals: A powerful tool to illuminate informative contents," *IEEE Reviews in Biomedical Engineering,* vol. 16, pp. 611 - 626, February 2022.

[3] M. Shaeri and A. M. Sodagar, "A framework for on-implant spike sorting based on salient feature selection," *Nature Communications,* vol. 11, no. 3278, pp. 1-9, 2020.

[4] A. M. Sodagar, G. E. Perlin, Y. Yao, K. Najafi and K. D. Wise, "An Implantable 64-Channel Wireless Microsystem for Single-Unit Neural Recording," *IEEE Journal of Solid-State Circuits,* vol. 44, no. 9, pp. 2591 - 2604, September 2009.

[5] J. Aziz, R. Genov, B. Bardakjian, M. Derchansky and P. Carlen, "256-Channel Integrated Neural Interface and Spatio-Temporal Signal Processor," *IEEE International Symposium on Circuits and Systems (ISCAS),* pp. 5075-5078, 2006.

[6] A. M. Sodagar, N. M. M. A. Shaeri and Y. Khazaei, "Implant-Embedded Neural Signal Processing Dedicated to High-Density, Brain-Implantable Microsystems," *Bioelectronic Medicine (Part of Springer Nature),* p. In preparation, 2023.

[7] E. Kandel, J. Koester, S. Mack and S. Siegelbaum, Principles of Neural Science, 5th edition, McGraw-Hill, October 2012, pp. Vol. BME-17, No. 3, 238-247.

[8] N. A. Steinmetz, C. Aydin, A. Lebedeva, M. Okun, M. Pachitariu, M. Bauza, M. Beau, J. Bhagat, C. Böhm, M. Broux, S. Chen, J. Colonell, R. J. Gardner and B. Karsh,

"Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings," *Science,* vol. 372, no. 6539, pp. 1-10, April 2021.

[9] T. Li, B. Sun, K. Xia, Q. Zeng, T. Wu and M. S. Humayun, "Design and Fabrication of A High-Density Flexible Microelectrode Array," *Proceedings of the 12th IEEE International Conference on Nano/Micro Engineered and Molecular Systems (NEMS),* pp. 299-302, 2017.

[10] E. Musk and Neuralink, "An integrated brain-machine interface platform with thousands of channels," *Journal of Medical Internet Research,* vol. 21, no. 10, p. e16194, Oct 2019.

[11] A. S. Herbawi, O. Christ, L. Kiessner, S. Mottaghi, U. G. Hofmann, O. Paul and P. Ruther, "CMOS Neural Probe With 1600 Close-Packed Recording Sites and 32 Analog Output Channels," *Journal of Microelectromechanical Systems (JMEMS),* vol. 27, no. 6, pp. 1023-1034, Dec. 2018.

[12] A. Obaid, M. E. Hanna, Y.-W. Wu, M. Kollo, R. Racz, M. R. Angle, J. Müller, N. Brackbill, W. Wray, F. Franke, E. J. Chichilnisky, A. Hierlemann, J. B. Ding, A. T. Schaefer and N. Melosh, "Massively parallel microwire arrays integrated with CMOS chips for neural recording," *Science Advances,* vol. 6, no. 12, pp. 1-10, Mar. 2020.

[13] W. Tedjo and T. Chen, "An Integrated Biosensor System With a High-Density Microelectrode Array for Real-Time Electrochemical Imaging," *IEEE Transactions on Biomedical Circuits and Systems,* vol. 14, no. 1, pp. 20-35, Feb. 2019.

[14] E. Rezayat, M. Dehaqani, K. Clark, Z. Bahmani, T. Moore and B. Noudoost, "Frontotemporal coordination predicts working memory performance and its local neural signatures," *Nature Communications,* vol. 12, no. 1103, p. 1–10, 2021.

[15] R. H. Olsson and K. D. Wise, "A Three-Dimensional Neural Recording Microsystem With Implantable Data Compression Circuitry," *IEEE Journal of Solid-State Circuits,* vol. 40, no. 12, pp. 2796 - 2804, December 2005.

[16] M. Shaeri, A. Afzal and M. Shoaran, “Challenges and Opportunities of Edge AI for Next-Generation Implantable BMIs,” *IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS),* pp. 190-193, 2022.

[17] K. D. Wise, J. B. Angell and A. Starr, “An Integrated-Circuit Approach to Extracellular Microelectrodes,” *IEEE Transaction on Biomedical Engineering,* Vols. BME-17, no. 3, pp. 238-247, July 1970.

[18] “https://www.fcc.gov/engineeringtechnology/policy-and-rules-division/general/radio-spectrum-allocation,” Radio spectrum allocation. [Online].

[19] I. H. Stevenson and K. P. Kording, “How advances in neural recording affect data analysis,” *Nature Neuroscience,* vol. 14, pp. 139-142, January 2011.

[20] M. Judy, M. S. Amir and R. Lotfi, “A nonlinear signal-specific ADC for efficient neural recording,” *Biomedical Circuits and Systems Conference (BioCAS),* pp. 17-20, 2010.

[21] Y. Khazaei and A. M. Sodagar, “Multi-channel ADC with improved bit rate and power consumption for electrocorticography systems,” *IEEE Biomedical Circuits and Systems Conference (BioCAS),* pp. 1-4, 2019.

[22] M. Shaeri and A. M. Sodagar, "A Method for Compression of Intra-Cortically-Recorded Neural Signals Dedicated to Implantable Brain-Machine Interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 23, no. 3, pp. 485-497, May 2015.

[23] N. Yazdani and A. M. Sodagar, "Reduction of spatial data redundancy in implantable multi-channel neural recording microsystems," *IEEE Biomedical Circuits and Systems Conference (BioCAS),* pp. 1-4, 2014.

[24] N. Yazdani, A. Rashidi and A. M. Sodagar, “A modified whitening transform for the reduction of spatial data redundancy in multichannel neural recording implants,” *International Journal of Circuit Theory and Applications,* vol. 46, p. 2283–2298, 2018.

[25] K. G. Oweiss, A. Mason, Y. Suhail, A. M. Kamboh and K. E. Thomson, "A Scalable Wavelet Transform VLSI Architecture for Real-Time Signal Processing in High-Density Intra-Cortical Implants," *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol. 54, no. 6, pp. 1266-1278, 2007.

[26] M. Namavar, R. Lotfi and A. M. Sodagar, “A 10-b, 330nW third-order predictive SAR ADC dedicated to neural recording brain implants,” *20th IEEE Interregional NEWCAS Conference (NEWCAS),* pp. 1-4, 2022.

[27] S. R. Nason, A. K. Vaskov, M. S. Willsey, E. J. Welle, H. An, P. P. Vu, A. J. Bullard, C. S. Nu, J. C. Kao, K. V. Shenoy, T. Jang, H. S. Kim, D. Blaauw, P. G. Patil and C. A. Chestek, “A low-power band of neuronal spiking activity dominated by local

single units improves the performance of brain–machine interfaces," *Nature Biomedical Engineering,* vol. 4, p. 973–983, 2020.

[28] S. Barati and A. M. Sodagar, "Adaptive Spike Detection Method Based on Capacitor Arrays Dedicated to Implantable Neural Recording Microsystems," *Annual International Conference of the IEEE Engineering in Medicine and Biology,* pp. 1-4, 2010.

[29] I. Obeid and P. D. Wolf, "Evaluation of spike-detection algorithms for a brain-machine interface application," *IEEE Transactions on Biomedical Engineering,* vol. 51, no. 6, p. 905–911, June 2004.

[30] Y. Yang and A. J. Mason, "Optimization of nonlinear energy operator based spike detection circuit for high density neural recordings," *IEEE International Symposium on Circuits and Systems (ISCAS),* pp. 1396-1399, 2014.

[31] M. A. Shaeri, A. M. Sodagar and H. Abrishami-Moghaddam, "A 64-channel neural signal processor/compressor based on Haar wavelet transform," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society,* pp. 6409-6412, 2011.

[32] S. Mirzaei, H. Hosseini-Nejad and A. M. Sodagar, "Spike Detection Technique Based on Spike Augmentation with Low Computational and Hardware Complexity," *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC),* pp. 894-897, 2020.

[33] A. M. Sodagar, K. D. Wise and K. Najafi, "A Neural Signal Processor for an Implantable Multi-Channel Cortical Recording Microsystem," *International Conference of the IEEE Engineering in Medicine and Biology Society,* pp. 5900-5903, 2006.

[34] B. Gosselin and M. Sawan, “An Ultra Low-Power CMOS Automatic Action Potential Detector,” *IEEE International Symposium on Circuits and Systems,* vol. 17, no. 4, pp. 346 - 353, 2009.

[35] S. Farsiani and A. M. Sodagar, “Intertwined-Pulse Modulation: Novel Approach for Compressive Data Telemetry,” *Scientific Reports,* vol. 12, no. 11966, pp. 1-12, 2022.

[36] H. Hosseini-Nejad, A. Jannesari and A. M. Sodagar, “Data reduction based on Walsh-Hadamard transform for implantable neural recording microsystems,” *IEEE Transactions on Biomedical Circuits and Systems,* vol. 8, no. 1, pp. 129 - 137, 2014.

[37] H. Hosseini-Nejad and A. M. Sodagar, "A 128-channel discrete cosine transform-based neural signal processor for implantable neural recording microsystems," *International Journal of Circuit Theory and Applications,* vol. 43, no. 4, p. 489–501, 2013.

[38] S. Farsiani and A. M. Sodagar, "Hardware and Power-Efficient Compression Technique Based on Discrete Tchebichef Transform for Neural Recording Microsystems," *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC),* pp. 3489-3492, 2020.

[39] S. Farsiani and A. M. Sodagar, "Compact agile Tchebycheff transform variant for temporal compression of neural signals on brain-implantable microsystems," *Elsevier Integration,* vol. 90, pp. 171-182, 2023.

[40] T. Wu, W. Zhao, H. Guo, H. H. Lim and Z. Yang, "A Streaming PCA VLSI Chip for Neural Data Compression," *IEEE Transactions on Biomedical Circuits and Systems,* vol. 11, no. 6, pp. 1290-1302, Dec. 2017.

[41] Q. Ma, L. Guo, S. M. A. Zeinolabedin and C. Mayr, "Ultra-low Power and Area-efficient Hardware Accelerator for Adaptive Neural Signal Compression," *IEEE Biomedical Circuits and Systems Conference (BioCAS),* pp. 1-4, 2021.

[42] F. Chen, A. P. Chandrakasan and V. Stojanović, "A signal-agnostic compressed sensing acquisition system for wireless and implantable sensors," *IEEE Custom Integrated Circuits Conference,* pp. 1-4, 2010.

[43] M. Shoaran, M. Kamal, C. Pollo, P. Vandergheynst and A. Schmid, "Compact low-power cortical recording architecture for compressive multichannel data acquisition," *IEEE Transactions on Biomedical Circuits and Systems,* vol. 8, no. 6, p. 857-870, 2014.

[44] X. Liu, M. Zhang, T. Xiong, A. G. Richardson, T. H. Lucas, P. S. Chin, R. Etienne-Cummings and T. D. Tran, "A Fully Integrated Wireless Compressed Sensing Neural Signal Acquisition System for Chronic Recording and Brain Machine Interface," *IEEE Transactions on Biomedical Circuits and Systems,* vol. 10, no. 4, pp. 874-883, Aug. 2016.

[45] S. Y. Park, J. Cho, K. Lee and E. Yoon, "Dynamic Power Reduction in Scalable Neural Recording Interface Using Spatiotemporal Correlation and Temporal Sparsity of

Neural Signals," *IEEE Journal of Solid-State Circuits,* vol. 53, no. 4, pp. 1102-1114, 2018.

[46] A. Cuevas-López, E. Pérez-Montoyo, V. J. López-Madrona, S. Canals and D. Moratal, "Low-Power Lossless Data Compression for Wireless Brain Electrophysiology," *Sensors,* vol. 22, no. 10, pp. 1-19, May 2022.

[47] T. Okazawa and I. Akita, "A Time-Domain Analog Spatial Compressed Sensing Encoder for Multi-Channel Neural Recording," *Sensors,* vol. 18, no. 184, pp. 1-21, January 2018.

[48] S. Schmale, B. Knoop, D. Peters-Drolshagen and S. Paul, "Structure reconstruction of correlated neural signals based on inpainting for brain monitoring," *IEEE Biomedical Circuits and Systems Conference (BioCAS),* pp. 1-4, 2015.

[49] Y. Khazaei, A. A. Shahkooh and A. M. Sodagar, "Spatial redundancy reduction in multi-channel implantable neural recording microsystems," *42nd Annual International*

Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 898-901, 2020.

[50] Y. Yang, S. Boling and A. J. Mason, "A hardware-efficient scalable spike sorting neural signal processor module for implantable high-channel-count brain machine interfaces," *IEEE Transactions on Biomedical Circuits and Systems,* vol. 11, no. 4, p. 743–754, 2017.

[51] N. Pérez-Prieto, Á. Rodríguez-Vázquez and M. Delgado-Restituto, "Spatial Encoding Techniques in Time-Multiplexed Neural Recording Front-Ends," *19th IEEE International New Circuits and Systems Conference (NEWCAS),* pp. 1-4, 2021.

[52] M. Zamani and A. Demosthenous, "Feature Extraction Using Extrema Sampling of Discrete Derivatives for Spike Sorting in Implantable Upper-Limb Neural Prostheses," *IEEE Transaction on Neural Systems and Rehabilitation Engineering,* vol. 22, no. 4, pp. 716 - 726, July 2014.

[53] C. O. Okreghe, M. Zamani and A. Demosthenous, “A Deep Neural Network-Based Spike Sorting With Improved Channel Selection and Artefact Removal,” *IEEE Access,* vol. 11, pp. 15131-15143, 2023.

[54] J. Sun, T. Li, T. Guo, Y. Li, C. Fu and Y. Liu, “Toward Ultra-large Scale Neural Spike Sorting with Distributed Sorting Channels and Unsupervised Training,” *IEEE International Symposium on Circuits and Systems,* pp. 3448-3452, 2022.

[55] D. Valencia and A. Alimohammad, “Neural Spike Sorting Using Binarized Neural Networks,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 29, pp. 206-214, 2021.

[56] C. Fu, T. Guo, Y. Li and Y. Liu, “Unsupervised Continuous Time Domain Spike Sorting for Large Scale Neural Processing Systems,” *IEEE Biomedical Circuits and Systems,* pp. 355-401, 2021.

[57] W. Lemaire, E. R. Koleibi, T. Omrani, M. Benhouria, K. Koua, C. Quesnel and L.-P. Gauthier, “Preliminary Results from a 49-Channel Neural Recording ASIC with Embedded Spike Compression in 28 nm CMOS,” *20th IEEE Interregional NEWCAS Conference,* pp. 285-289, 2022.

[58] V. Karkare, S. Gibson and D. Marković, “A 75-µW, 16-Channel Neural Spike-Sorting Processor with Unsupervised Clustering,” *IEEE Journal of Solid-State Circuits,* vol. 48, no. 9, pp. 2230 - 2238, 2013.

[59] F. Kalantari, H. Hosseini-Nejad and A. M. Sodagar, “Hardware-Efficient, On-the-Fly, On-Implant Spike Sorter Dedicated to Brain-Implantable Microsystems,” *IEEE Transactions on Very-Large-Scale Integration,* vol. 30, no. 8, pp. 1098-1106, August 2022.

[60] A. V. Oppenheim and e. al, Signals and Systems, 2nd edition, Pearson, 1996.

[61] D. R. I. Nauhaus, “CRCNS - Collaborative Research in Computational Neuroscience - Data sharing,” Single- and multi-unit recordings from monkey primary visual cortex., 2009. [Online]. Available: http://crcns.org/data-sets/vc/pvc-1.

[62] L. M. McGuire, G. Telian, K. Laboy-Juárez, T. Miyashita, D. Lee, K. A. Smith and D. E. Feldman, “CRCNS - Collaborative Research in Computational Neuroscience - Data sharing,” Extracellular recordings in rat barrel cortex during a whisker based discrimination task of tactile stimuli that varied in whisker deflection speed at short and fast time scales, 2016. [Online]. Available: https://crcns.org/data-sets/ssc/ssc-4.

[63] M. Wehr and H. Asari, “CRCNS - Collaborative Research in Computational Neuroscience - Data sharing,” Neuronal responses to various natural and synthetic sounds were recorded using whole-cell and cell-attached recording techniques in the primary auditory cortex and auditory thalamus (medial geniculate body; MGB) in the anesthetized rat, 2002-2007. [Online]. Available: https://crcns.org/data-sets/ac/ac-1.

[64] M. Nekoui and A. M. Sodagar, "Spike Compression through Selective Downsampling and Piecewise Curve Fitting Dedicated to Neural Recording Brain Implants," *IEEE Biomedical Circuits and Systems Conference (BioCAS),* pp. 50-54, 2022.

[65] M. R. Spiegel, Probability and Statistics, Mc Graw Hill, 2009.

[66] P. B. Patnaik, "The Non-Central X2-and F-Distribution and their Applications," *Biometrika,* vol. 36, pp. 202-232, 1949.

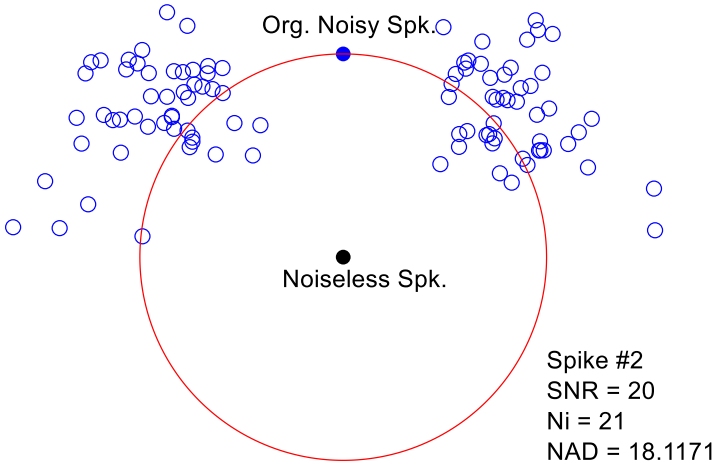
[67] J. D. J. W. C. H. A. H. E. T.-K. M. B. SN Flesher, "A brain computer interface that evokes tactile sensations improves robotic arm control," *Science,* vol. 372, pp. 831-836, 2021.

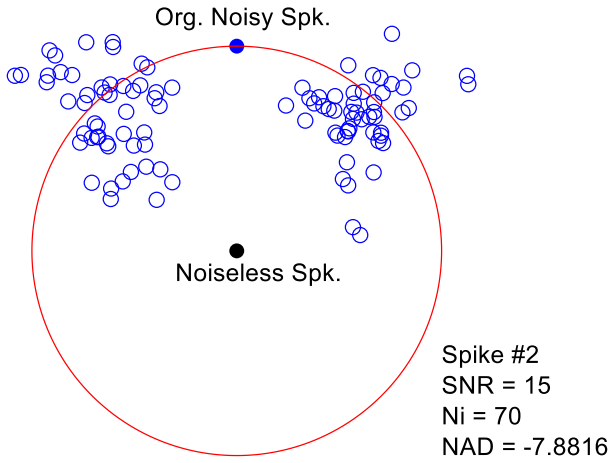
[68] F. R. Willett, D. T. Avansino, L. R. Hochberg, J. M. Henderson and K. V. Shenoy, "High-performance brain-to-text communication via handwriting," *Nature,* vol. 593, pp. 249-254, 2021.

[69] "Terasic Inc.," FPGA Board, 24 Febuary 2022. [Online]. Available: https://www.terasic.com.tw/en/.

Appendix A Denoising Property

In order to examine the denoising effectiveness of the spike compression approach proposed in this study, distinct sets of 100 pre-recorded extracellular neural spikes are organized for each waveshape, as depicted in Figure 2-7. These sets encompass a variety of SNRs. The SNR spans from 5dB to 25dB for each individual spike waveshape. The outcomes of this investigation, shown in Figure A-1~6, are visually represented in through the dissimilitude diagrams. These diagrams portray the study's findings in associated with the denoising capability of the proposed compression approach.


Org. Noisy Spk.
Noiseless Spk.
Spike #2
SNR = 20
Ni = 21
NAD = 18.1171


Org. Noisy Spk.
Noiseless Spk.
Spike #2
SNR = 15
Ni = 70
NAD = -7.8816

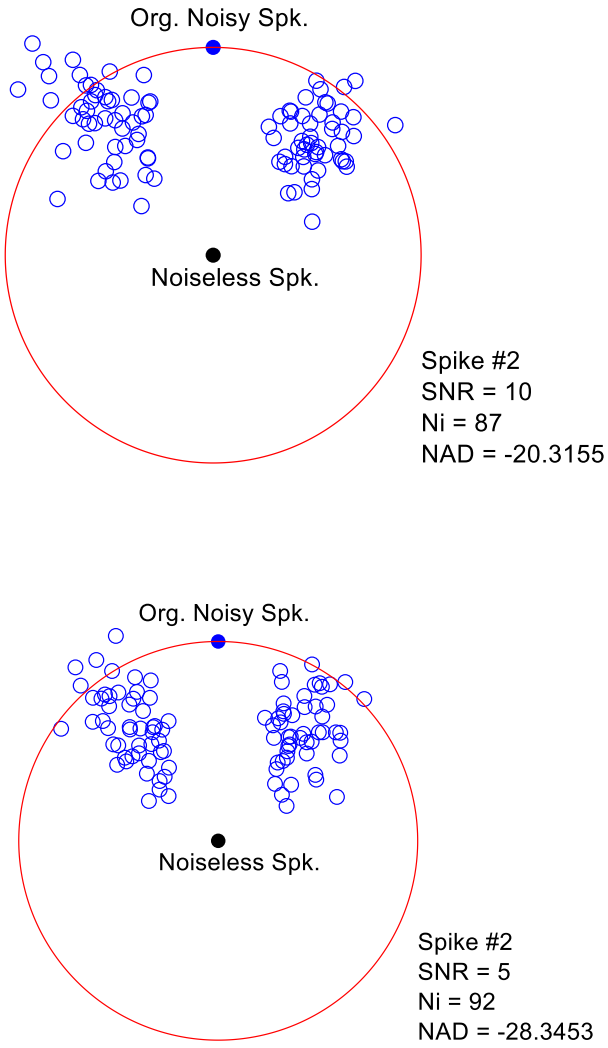
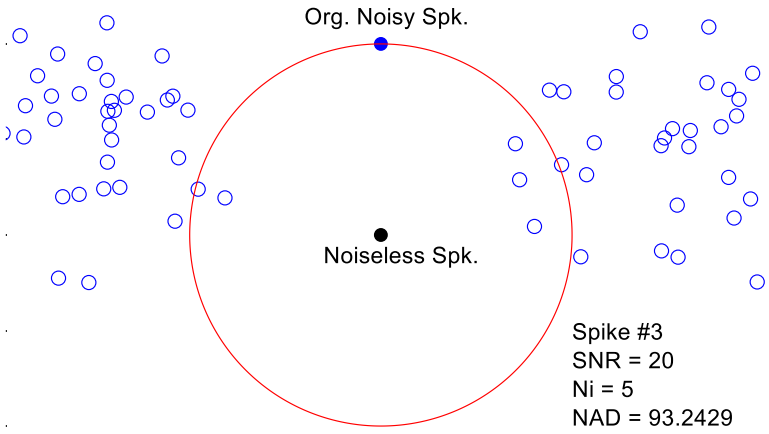
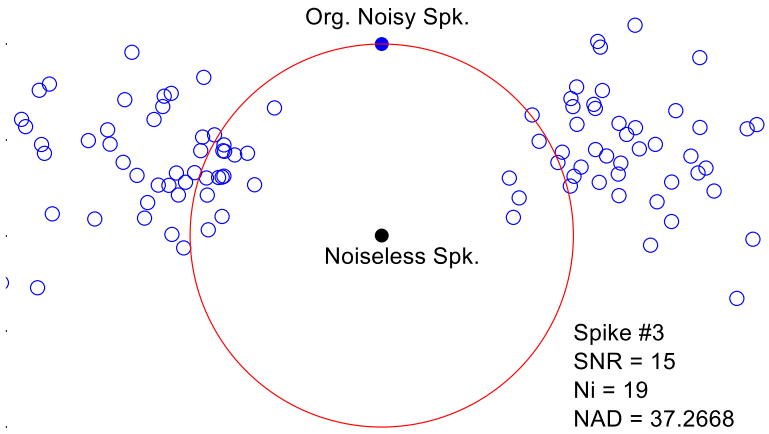


Figure A-0-1: The denoising capability of the proposed spike compression approach. The dissimilitude diagram for a class of spikes with waveshape #2 (according to Figure 2-7) for SNR=20, 15, 10, and 5dB.


Org. Noisy Spk.
Noiseless Spk.
Spike #3
SNR = 20
Ni = 5
NAD = 93.2429


Org. Noisy Spk.
Noiseless Spk.
Spike #3
SNR = 15
Ni = 19
NAD = 37.2668

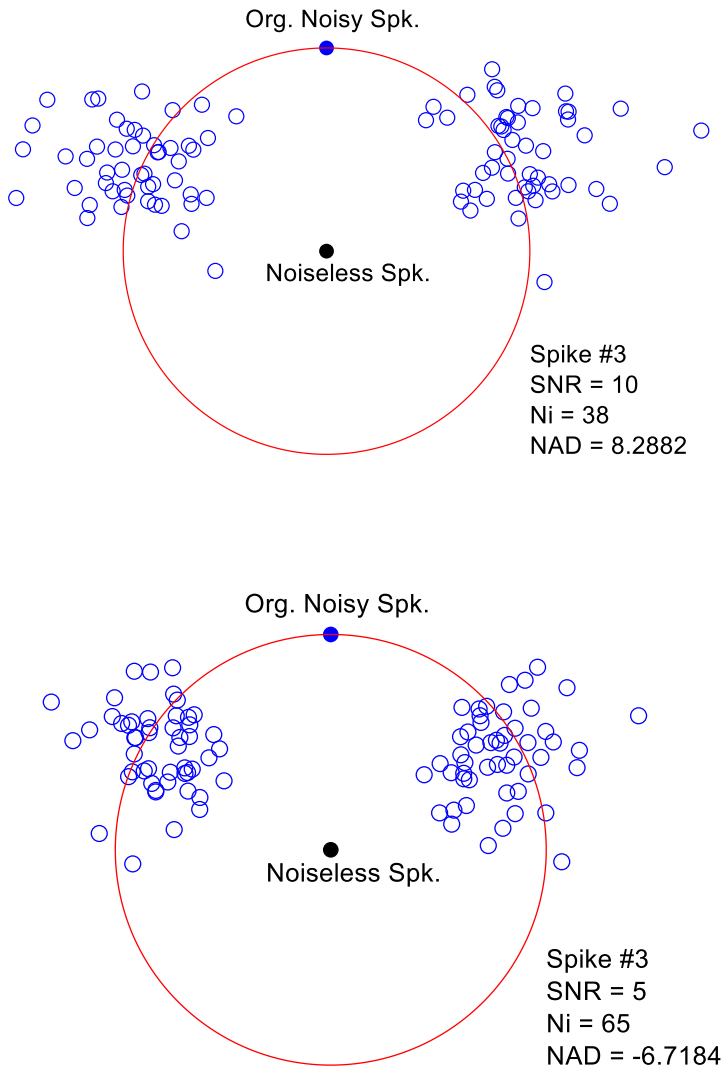
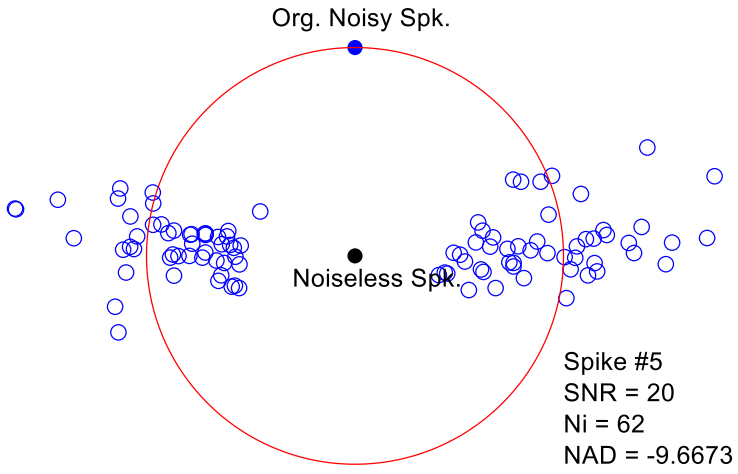
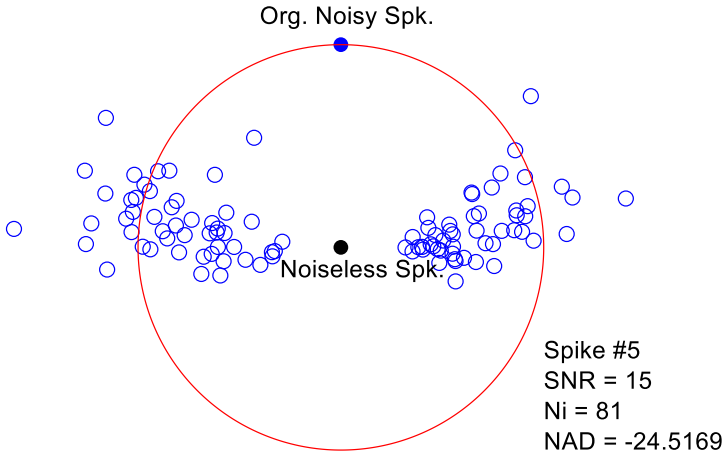


Figure A-2: The denoising capability of the proposed spike compression approach. The dissimilitude diagram for a class of spikes with waveshape #3 (according to Figure 2-7) for SNR=20, 15, 10, and 5dB.


Org. Noisy Spk.
Noiseless Spk.
Spike #5
SNR = 20
Ni = 62
NAD = -9.6673


Org. Noisy Spk.
Noiseless Spk.
Spike #5
SNR = 15
Ni = 81
NAD = -24.5169

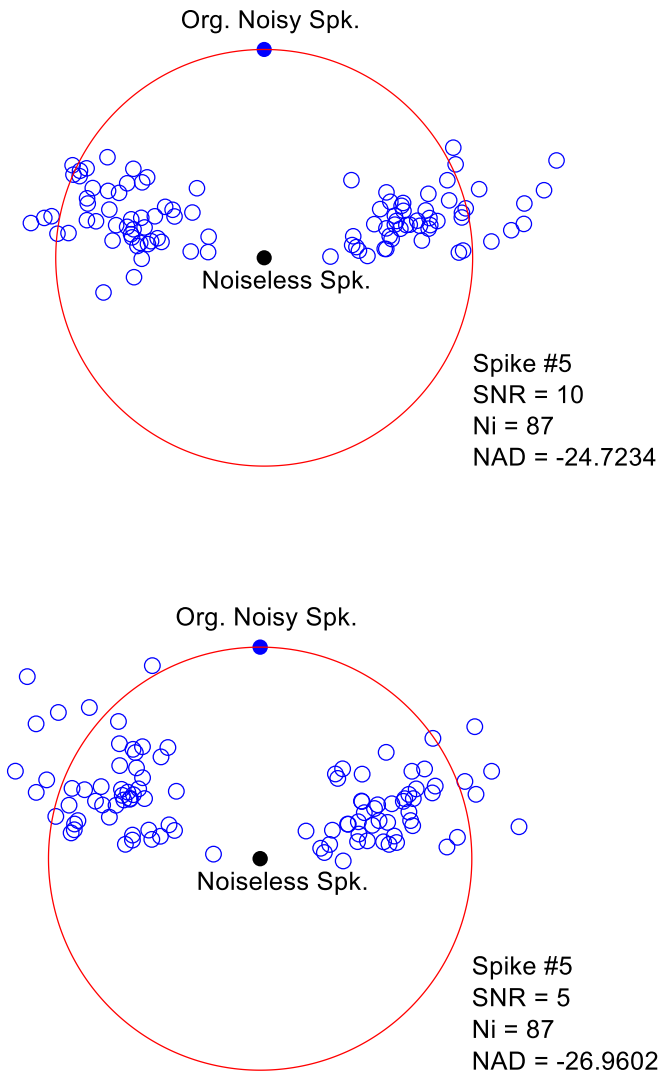
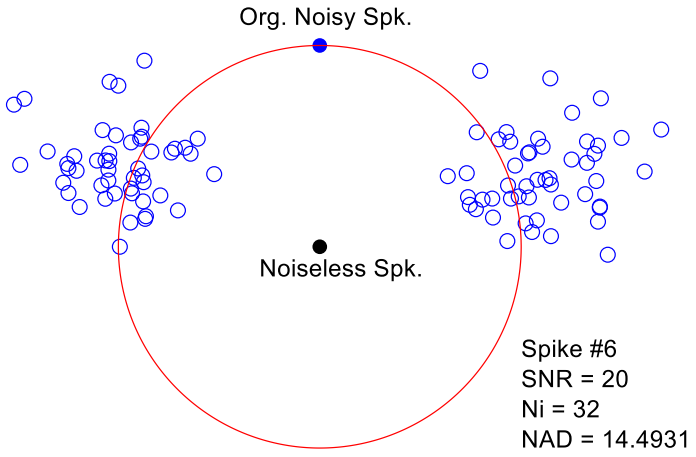
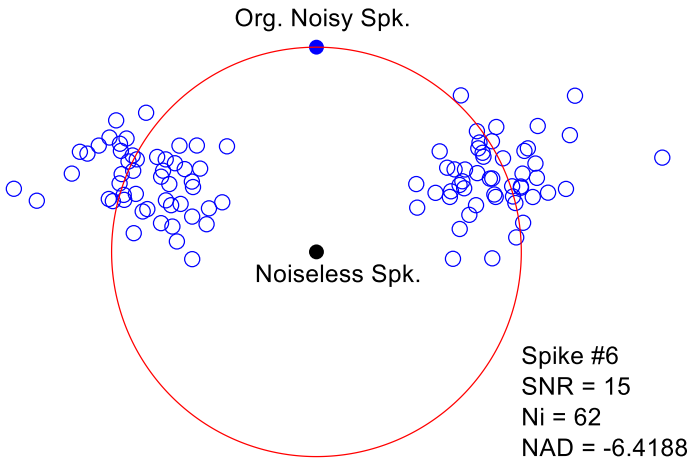


Figure A-3: The denoising capability of the proposed spike compression approach. The dissimilitude diagram for a class of spikes with waveshape #5 (according to Figure 2-7) for SNR=20, 15, 10, and 5dB.


Org. Noisy Spk.
Noiseless Spk.
Spike #6
SNR = 20
Ni = 32
NAD = 14.4931


Org. Noisy Spk.
Noiseless Spk.
Spike #6
SNR = 15
Ni = 62
NAD = -6.4188

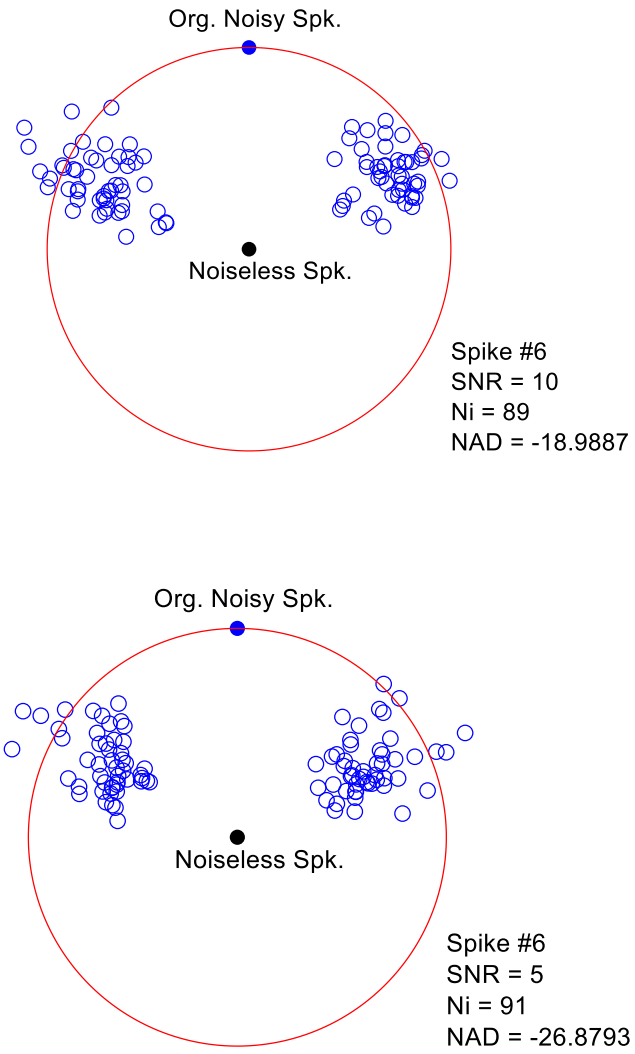
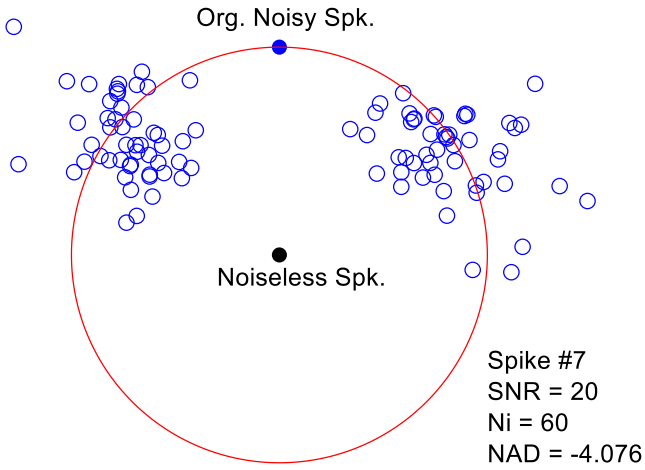
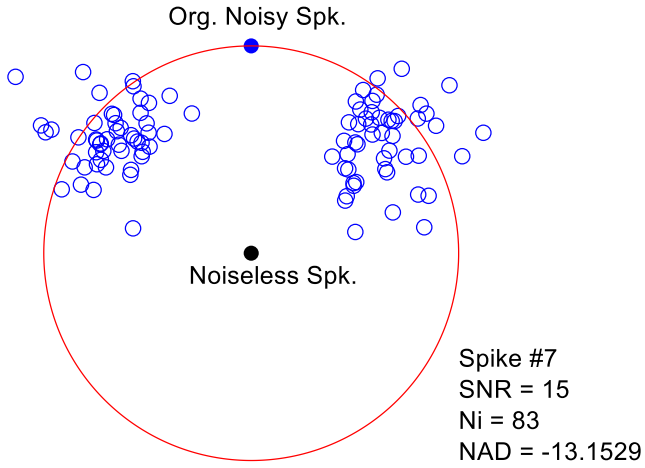


Figure A-4: The denoising capability of the proposed spike compression approach. The dissimilitude diagram for a class of spikes with waveshape #6 (according to Figure 2-7) for SNR=20, 15, 10, and 5dB.


Org. Noisy Spk.
Noiseless Spk.
Spike #7
SNR = 20
Ni = 60
NAD = -4.076


Org. Noisy Spk.
Noiseless Spk.
Spike #7
SNR = 15
Ni = 83
NAD = -13.1529

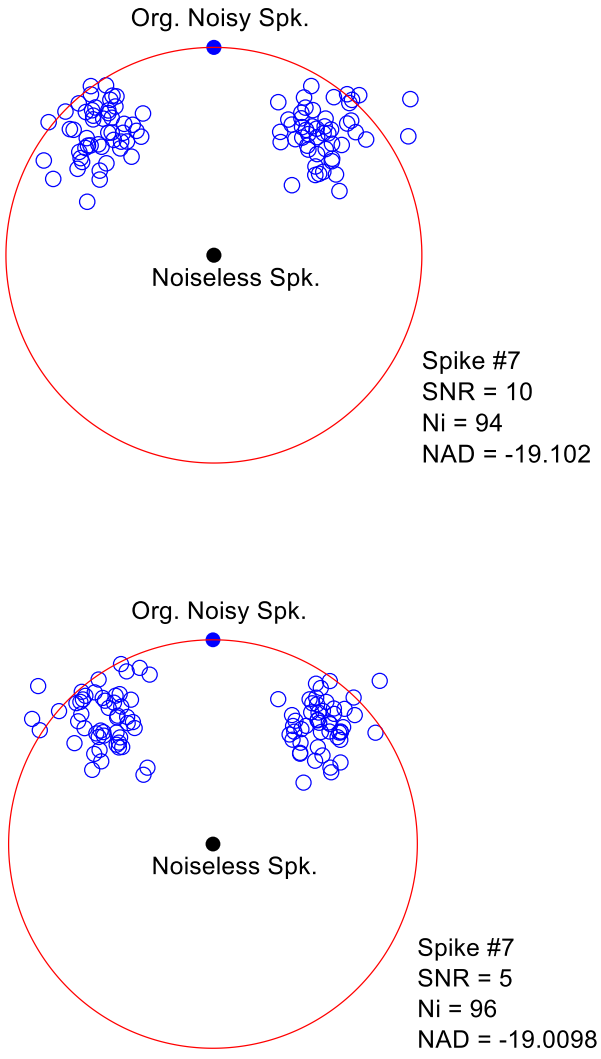
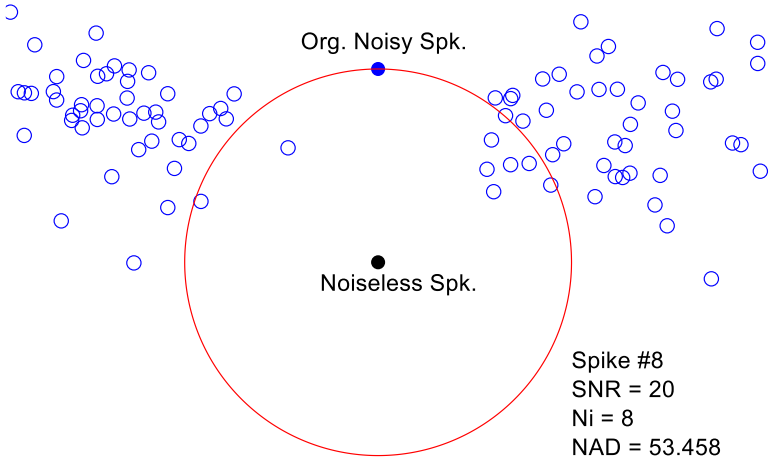
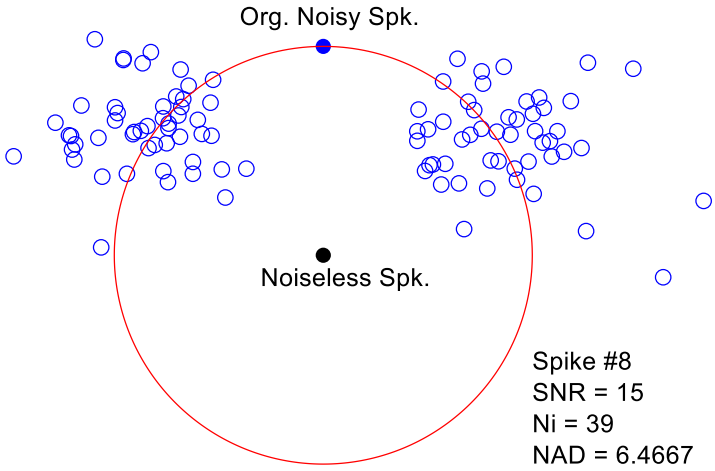


Figure A-5: The denoising capability of the proposed spike compression approach. The dissimilitude diagram for a class of spikes with waveshape #7 (according to Figure 2-7) for SNR=20, 15, 10, and 5dB.


Org. Noisy Spk.
Noiseless Spk.
Spike #8
SNR = 20
Ni = 8
NAD = 53.458


Org. Noisy Spk.
Noiseless Spk.
Spike #8
SNR = 15
Ni = 39
NAD = 6.4667

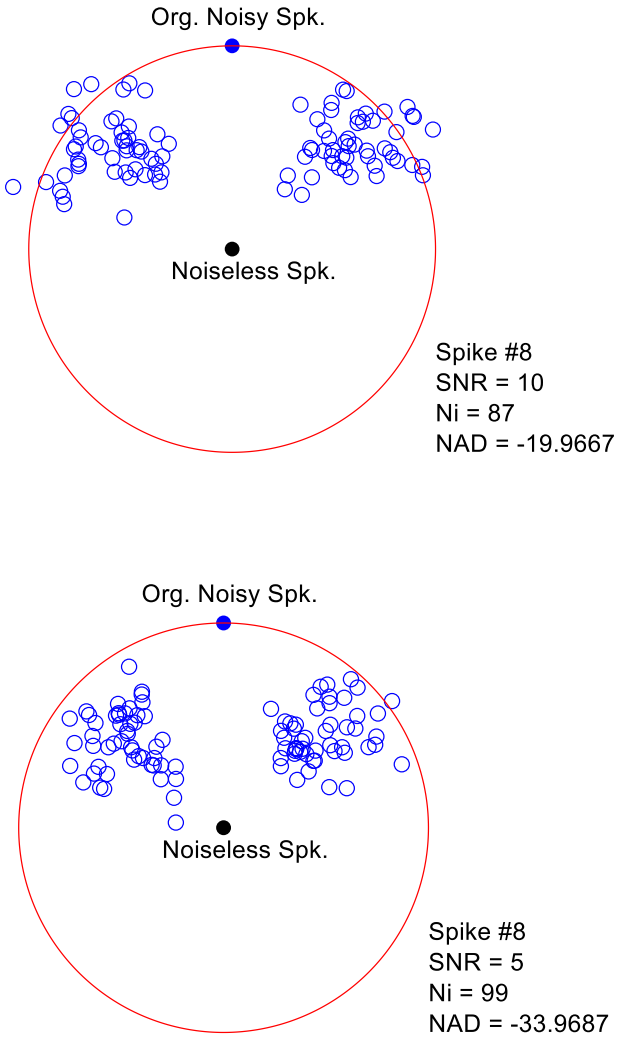


Figure A-6: The denoising capability of the proposed spike compression approach. The dissimilitude diagram for a class of spikes with waveshape #8 (according to Figure 2-7) for SNR=20, 15, 10, and 5dB.

Appendix B Dataset Information

B.1 Information about pvc-1

Data Collection: The data presented here represent the responses of V1 populations of neurons to natural image sequences. Data was obtained using 10 x 10 extracellular electrode arrays. Normally about 50 electrodes provided reliable responses. A description of surgical procedures, anaesthesia, and the insertion of the arrays can be found in: Precise Alignment of Micromachined Electrode Arrays with V1 Functional Maps. Ian Nauhaus and Dario L. Ringach. J Neurophysiol. 2007 May;97(5):3781-9. (Pubmed ID 17344376)

All recordings were performed on Old-world monkeys (Macaca fascicularis) in primary visual cortex (area V1). Receptive fields were between 2 and 6 degrees in eccentricity.

Visual stimulus: Natural image sequences were generated by digitally sampling commercially available videotapes in VHS/NTSC format. A Silicon Graphics R10000 Solid Impact was used to sample frames at a spatial resolution of 320 × 240 pixels and at a temporal rate of 30 Hz. Thirty different segments of approximately 30-s duration were sampled from four

different movies (Sleeper, Benji, Goldfinger and Sheakespere in Love), making a total of about 60 minutes of video.

B.2 Information about ssc-4

The data set described in this document was collected from extracellular spike recordings in whisker primary somatosensory cortex (S1) of rats performing a whisker-based discrimination task. Neural activity is described by the spike times of isolated single and multiunit clusters obtained from spike sorting from moveable tetrode arrays. The whisker stimuli consisted of rapid whisker impulse sequences. Each sequence (120-150 ms total duration) contained 3 brief impulses (16-26 ms each) with either Fast (F), Medium (M) or Slow (S) rise-fall velocity. Sequences had FFF, FMS, SMF or SSS order. Thus, sequences varied in whisker deflection speed at short (5-20 ms) and long (150 ms) time scales. On each trial, one stimulus was delivered, and rats discriminated stimuli in a 2-alternative forced choice task. Hence, this data set serves as a resource to explore the time scales at which cortical neurons encode whisker stimuli and how their spiking activity correlates to sensory perception. The structure contains trial information and complete neurophysiology data from 5 rats, over a total of 80 recording sessions. Neural recordings are from layer 2/3 to 6 of S1, and include both fast-spiking and regular spiking single units (identified separately).

B.3 Information about ac-1

The data consists of electrophysiology data recorded by Michael Wehr (from 2002 to 2003) and Hiroki Asari (from 2005 to 2007) at Anthony Zador lab at Cold Spring Harbor Laboratory. Neuronal responses to various natural and synthetic sounds were recorded using whole-cell and cell-attached recording techniques in the primary auditory cortex (area A1) and auditory thalamus (medial geniculate body; MGB) in the anesthetized rat.