33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Feasibility Study for Future Implantable Neural-Silicon Interface Devices

Allann Al-armaghany, Bo Yu, Terrence Mak, Kin-Fai Tong and Yihe Sun

Abstract—**The emerging neural-silicon interface devices bridge nerve systems with artificial systems and play a key role in neuro-prostheses and neuro-rehabilitation applications. Integrating neural signal collection, processing and transmission on a single device will make clinical applications more practical and feasible. This paper focuses on the wireless antenna part and real-time neural signal analysis part of implantable brain-machine interface (BMI) devices. We propose to use millimeter-wave for wireless connections between different areas of a brain. Various antenna, including microstrip patch, monopole antenna and substrate integrated waveguide antenna are considered for the intra-cortical proximity communication. A Hebbian eigenfilter based method is proposed for multi-channel neuronal spike sorting. Folding and parallel design techniques are employed to explore various structures and make a trade-off between area and power consumption. Field programmable logic arrays (FPGAs) are used to evaluate various structures.**

I. Introduction

There has been much interest in developing brain-machine interface (BMI) technology to help improve the quality of life, to restore motor or cognitive functions for people with severe disabilities and injuries [1]. New insights and discoveries into accessing cortical circuits through various real-time closed-loop implantable micro-systems equipped with multi-electrode arrays (MEAs) are revealed. These implantable systems typically allow neural activity to control external prostheses or devices [2]

Recently, the BMI approach has been extended to create a recurrent brain-machine-interface [1] system in which neural activity is recorded and processed in real-time to generate electrical stimulation to particular areas of the brain through implantable electrodes. Such recurrent interface introduces an artificial neural pathway that the adaptive brain could learn to adapt and incorporate into normal desired function. This autonomous recurrent BMI paradigm opens new experimental directions and opportunities for promising clinical applications. It is at the same time presents novel design challenges at both the circuit and system levels due to the severe power and size constrains, biological interface restrictions and real-time processing requirements.

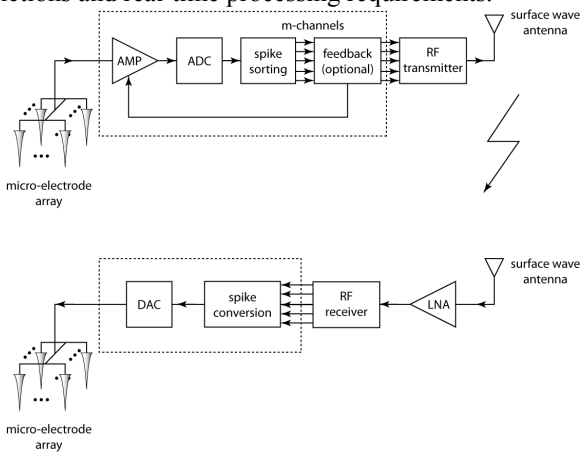


Fig. 1 System diagram of the proposed wireless system.

We propose a closed-loop neural signal recording and stimulus system for recurrent BMI paradigm, as shown in Fig. 1. The recording part consists of analog amplifier, analog-to-digital converter, neural signal processing hardware, wireless transmitter circuits and antenna. The stimulus side receives data and control signal through antenna and wireless receiver. The stimulus signals are obtained through signal processing and digital-to-analog conversion and finally applied on micro-electrodes.

Implementing artificial neural pathways across different areas of the brain requires interconnections of communicating between multiple separate recording and stimulating interfaces. The design of such communication channel becomes the key to the overall micro-system performance and which also determines the overall power dissipation and hardware area. To communicate wirelessly provides an appealing solution to the biological restriction of implantable devices. Issues such as forming scars or chance of infection could be therefore minimized.

We argue that for such recurrent BMI applications, design constrains and communication criteria are rather different, such as pre-determined and fixed receiver location and communicating in proximity in the brain regions. These provide an opportunity to optimize the communication channels in the artificial neural pathways in order to improve the sustainability and efficiency of the device. In this paper, we present three micro-antenna designs that can be readily applied for the wireless proximity communication channels. These new designs exploit the surface wave propagations at

This work was supported by EPSRC Knowledge Transfer Account.

The first and the second author have the same contribution.

Allann Al-armaghancy is with the Department of Electronic and Electrical Engineering, UCL (e-mail: allann.al-armaghany@ee.ucl.ac.uk).

Bo Yu and Yihe Sun are with Tsinghua National Laboratory for Information Science and Technology, Institute of Microelectronics, Tsinghua University, Beijing 100084, China.

Kin-Fai Tong is with the Department of Electronic and Electrical Engineering, UCL, Torrington Place, WC1E 7JE, UK (phone: +44 (0)20-7679-3019; fax: +44 (0)20-7388-9325; e-mail: k.tong@ee.ucl.ac.uk).

Terrenc Mak is with School of Electrical, Electronic and Computer Engineering, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK. His is also with the Institute of Neuroscience, Newcastle Biomedicine, at the same University (e-mail: terrence.mak@ncl.ac.uk)

60 GHz and can significantly reduce hardware area and power dissipation for the necessary communications.

Because no external computing engine is involved, on-site real-time neural signal processing is required by a close-loop BMI system. Spike sorting [6] that detects neuronal spikes from noises and discriminates different neuronal spikes should be employed to turn a raw neural recording into separate spike trains. However, extremely large data bandwidth and volume associated with multi-channel recording requires high computational effort to achieve real-time spike sorting and presents a major design challenge in terms of power and area consumption.

In general, software solutions are difficult to meet the real-time requirement, power and area constraints due to lack of parallel and specialized computing structure. Parallel hardware structures are preferred for multi-channel real-time neural signal analysis. We propose a Hebbian eigenfilter [7] based spike sorting algorithm, which is computational efficient and suitable for low-cost hardware implementation. Based on the proposed algorithm, a multi-channel spike sorting hardware structure is designed. Parallel computing structures are used to achieve real-time multi-channel processing. Folding techniques are employed to share computing resources among channels and reduce area overhead.

II. DESIGN

A. Antennas

Addressing on the size and power constraints of implanted systems, we propose to use a 60 GHz wireless system to realize the artificial neural pathways. Physical small antennas are desirable for embedding antenna systems under the skin of heads. Common size reduction techniques for conformal antennas include using high-dielectric-constant materials as substrate [3] or cutting slots on the antennas for extending current path lengths [4], however these techniques cannot support the highly miniature size requirement for implanted systems. The advantages of using 60 GHz systems for implantable systems are:

i) Inherent compact size: The physical size of an antenna basically depends on the wavelength of the operating frequency. The free space wavelength at 60 GHz is 5 mm; therefore without using any complex size-reduction techniques; simple structured and highly efficient small antenna are possible.

ii) Short-range communication: It was commonly believe that the propagation loss of electromagnetic wave at 60 GHz is significant, so it is not suitable for practical wireless systems. However, for short-range communications, like this project, it is considered to be strength rather than weakness, as the interference from other systems is negligible.

iii) RF modules availability: Recently, the designs of monolithic RF modules for 60 GHz systems are becoming mature. Reliable RF modules are available for the feasibility investigation. Resources can be focused on the antenna design and module integrations.

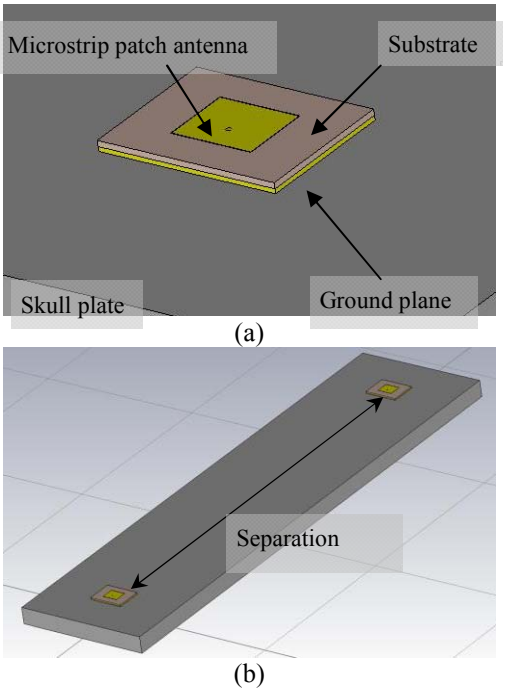


Fig. 2 Geometry of (a) single microstrip patch antenna and (b) the antenna system geometry.

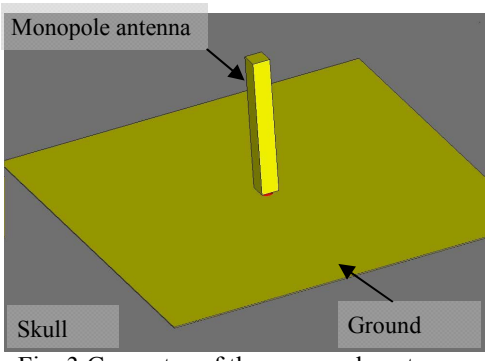


Fig. 3 Geometry of the monopole antenna.

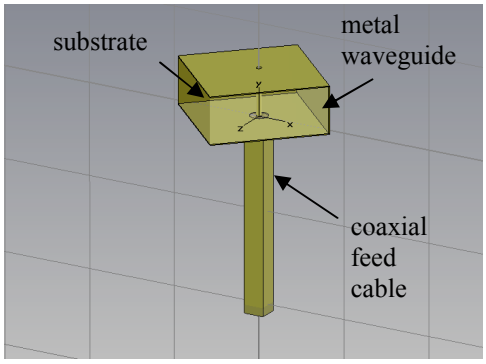


Fig. 4 Geometry of an SIW antenna.

We have investigated three different types of antennas, namely microstrip patch antennas, monopole antennas and substrate integrated waveguide antennas. The dimensions of the investigated antennas are within $5 \times 5 \times 1.5$ mm^3. This size will satisfy the compact size requirement for implanted systems. Commercial EM software CST Microwave Studio 2010 [5] has been used in the simulations.

1) Microstrip Patch Antennas

Microstrip patch antennas are thin, and can be easily fabricated by low cost printed circuit technique. Because of its thickness, it is suitable for mounting on curved surfaces. It consists of a conductive patch as the radiating element on the top surface of a thin dielectric substrate separating them from the ground plane on the bottom. A patch is typically wider than a strip and its shape and dimensions are important features of the antenna. Two microstrip patch antennas are designed on two pieces of microwave substrate (thickness = 0.127 mm and dielectric constant r = 2.2) with ground plane support as shown in Fig. 2 (a) and (b).

2) Monopole Antennas

A monopole antenna consists of a vertical straight wire locating on a flat ground plane as shown in Fig. 3. In contrast to microstrip patch antennas, monopole antennas radiates sideway, rather than at the boresight. This radiation characteristic enhances the signal strength by better design of the radiation direction.

3) Substrate Integrated Waveguide Antennas (SIW)

A SIW antenna, as described by its name, is basically a substrate filled metallic waveguide. However, the SIW antenna studied in this paper, the two openings are adjacent to, rather than opposite to each other as in conventional waveguides. The SIW antenna is also probe-fed by a 50 ohm coaxial cable as shown in Fig. 4.

B. Multi-channel spike sorting

1) Hebbian eigenfilter based spike sorting

Usually neural spike trains recorded by a single electrode contain various noises and neuronal spikes from more than one neuron. Noises and superimposed neuronal spikes in the spike trains compromise the performance and accuracy of the system. Spike sorting is a kind of process to detect neuronal spikes from noises and discriminate different neuronal spikes to corresponding neurons.

In general, a spike sorting algorithm consists of spike detection, spike alignment, spike feature extraction and clustering. A commonly used detection method is to compare absolute voltage of the recorded signal with a threshold that is derived from median value of the raw spike train. However, hardware cost of obtaining the median value is high. Another method is the nonlinear energy operator (NEO) based spike detection. It has not only small hardware cost but also high accuracy by considering spike amplitude and frequency [8].

Feature extraction projects spikes into the feature space that highlights differences in the spikes. Principal component analysis and wavelet transformation are two basic feature extraction algorithms. However, computational expensive nature makes them not suit for on-chip hardware implementation. We have proposed to use general Hebbian algorithm (GHA) [9] to accomplish PCA based feature extraction. GHA has several advantages in terms of hardware implementation. First, it does not involve covariance matrix computation and the following eigenvector calculation of covariance matrix or single value decomposition. Second, it has the ability to filter specified number of most important PCs. This filtering nature is very useful, because in most case the first few PCs capture most variances of aligned spikes, only computing the leading PCs save lots of computation effort.

After feature extraction, k-means clustering algorithm is employed to calculate the centroid of each cluster in the feature space. According to the distance to centroids, spikes in the feature space are assigned to its nearest centroid. Because of computational simplicity, k-means algorithm quite suits for on-chip implementation.

Fig. 5 summarizes our spike sorting algorithm. NEO based method, Hebbian eigenfilter and k-means algorithm are employed to train a threshold, principal components and centroids for on-line spike sorting. On-line spike sorting detects spikes, projects spikes to feature space and performs clustering in real time.

2) Hardware structure

The most intuitive structure supporting real-time multi-channel spike sorting is a fully parallel structure that allocates each channel independent on-line processing hardware and training hardware. However, this fully parallel structure has extremely high hardware area overhead and does not suit for systems with hundreds and thousands of channels.

Folding technique is utilized to share computing hardware among multi-channels. Fig. 6 shows the structure of M-channel N-folded spike sorting hardware. It consists of two parts, the parameter training hardware and real-time processing hardware. The real-time processing hardware performs M channels real-time spike sorting through M/N N-folded processing cores. Parameter training hardware implements Hebbian eigenfilter, k-means and threshold estimation algorithms. Since the performance requirement of training hardware is relatively low (4.32 second per channel for a 10,000 electrodes system) [10], all real-time processing hardware shares the same training hardware.

In real-time processing part, NEO based spike detection and peak alignment work in cycle basis. Comparator compares each NEO filtered sample with a threshold. Once one sample crosses the threshold and a following peak is detected, the peak detection asserts a spike event to the controller and aligned spikes buffered in the fifo are sent to real-time projection hardware. The real-time projection hardware works in event basis and is enabled by the controller after a spike event is detected. Multiply- accumulate units, memories storing trained principal com- ponents and score registers are the main parts of the real-time projection hardware. Dot production between spikes and trained principal components are conducted by multiply- accumulate units to project spikes to feature space. The intermediate results of dot production are stored in feature score registers. After real-time projection finishes, the norm units calculate the distances between final projection score and each trained centroids. The minimizer assigns a score to its nearest centroid according to the norm results.

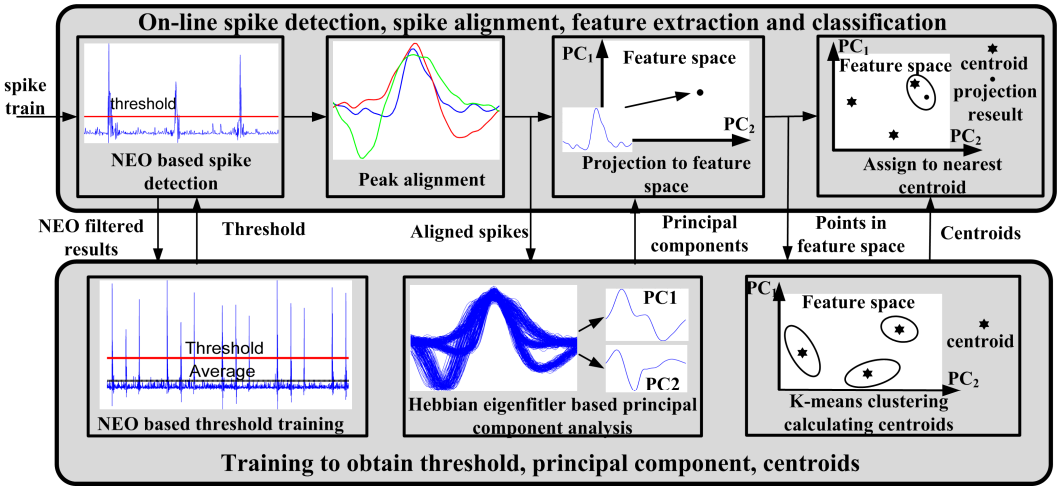


Fig. 5 Hebbian eigenfilter based spike sorting.

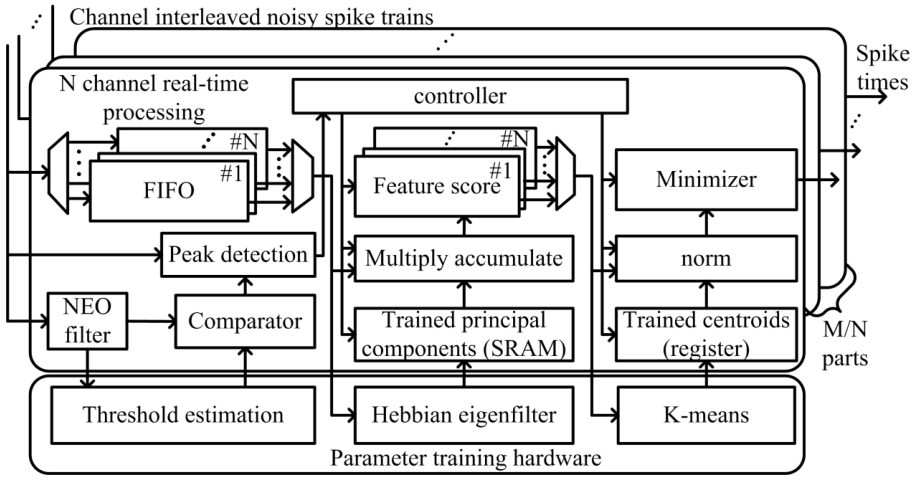


Fig. 6 Hardware structure of multi-channel spike sorting.

III. Evaluation Results

A. Antenna

1) Microstrip Patch Antennas

S-Parameters: Fig. 7 shows the S11 as a function of frequency when the two antennas are 40 mm apart. We can see that the operating bandwidth covers from 58.9 to 61.1 GHz (2.2 GHz absolute bandwidth), and centers at about 60 GHz. The bandwidths of the antennas do not vary significantly between the investigated separations.

The path losses of the system at different separations are depicted in Fig. 8. We can observe that the loss is about 55dB when the separation is 40mm and it increases about 10dB for every 10 mm extension. Considering a low power system, if the power at the output of the transmitting antenna is 1 μW or -30dBm, then the power at the input of the receiving antenna will be -85dBm, which is a receivable level of most conventional millimeter-wave receivers.

Far field radiation: From Fig. 9, the maximum gain of the antenna system is 7.76dBi at +z direction, while the gain is -10dBi at the direction towards the transmitting or receiving antenna. The -85dBm signal discussed in the previous section can be improved by about 35dB if the antenna main lobes are designed to radiation at the desired direction. This also implies that the communication range can be extended.

2) Monopole Antennas

S-Parameters: The S11 as a function of frequency is shown in Fig. 10. We can see that the operating bandwidth (S11 < -10dB) covers from 54 to 68 GHz (14 GHz absolute bandwidth), and centers at about 60 GHz. As shown in Fig. 10, the S21 or isolation between the transmitting and receiving antennas is approximately -42dB across the operating band when the separation between the two monopole antennas is 40 mm. When compared to that of microstrip patch antennas system (the first case), the improvement is 13 dB. It demonstrates that specific designed antennas can significantly improve the system performance.

Far field radiation: In Fig. 11, the maximum gain of the antenna is only 3.7 dBi when is about 70°, while the gain is 0 dBi at the desired direction (90°). Moreover, it can be observed that the radiation pattern is somehow omni-directional, it radiates almost equally in the xy-plane, therefore, more directional antenna design can further focus the radiation beam to particular desired direction for even higher S21.

Surface wave propagation: The normal component of the electric field travels along the skull model plate is shown in Fig. 12. We can observe that the wave propagates radially, and reflects at the edges of the model causing interferences. This creeping mode of propagation will be useful for non

line-of-sight communication, such as an implant system on the skull surface we are proposing in this paper.

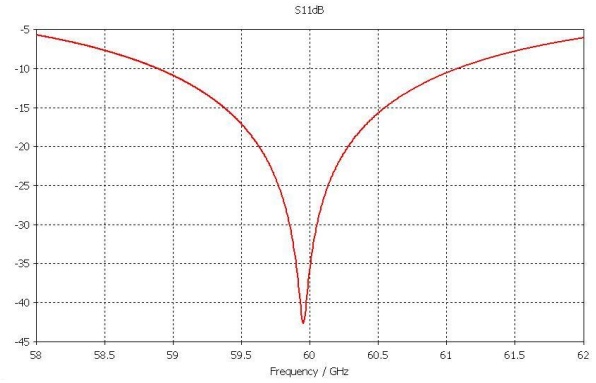


Fig. 7 Simulated S11 of the microstrip patch antenna.

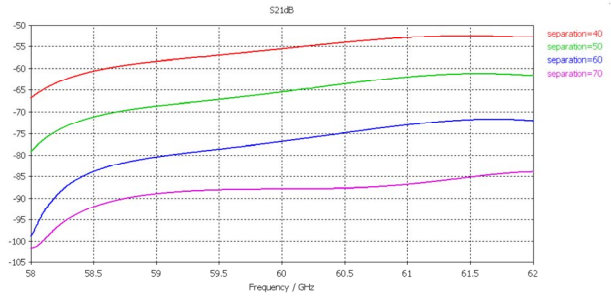
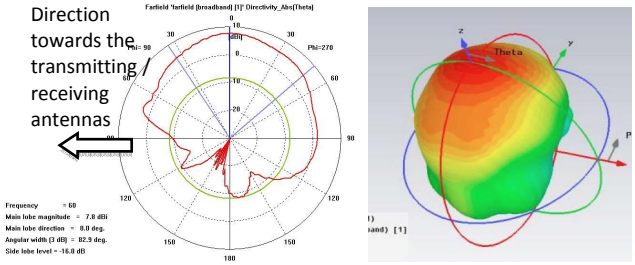


Fig. 8 The path loss of the microstrip patch antenna system at different separations.



(a) (b)

Fig. 9 (a) E-plane and (b) 3-D radiation patterns of the microstrip patch antenna at 60 GHz.

3)Substrate Integrated Waveguide Antennas (SIW)

S-Parameters: The impedance bandwidth of the SIW antenna shown in Fig. 13 covers 58.6 to 61.4 GHz (2.8 GHz absolute bandwidth), which is compatible to that of the microstrip patch antenna. The S21 information was not available yet when preparing the paper, because of the huge size of the model at the millimeter wave frequency band. But it will be provided in the presentation.

Far field radiation: As illustrated in Fig. 14, the SIW antenna has more directional radiation pattern when compared to the previous two cases. Moreover, as the antenna is mostly enclosed by the metal waveguide, we believe that it is more robust to the change of surrounding environment. A newer SIW antenna geometry is being investigated for further diverting the lower half of the radiation pattern to the upper half.

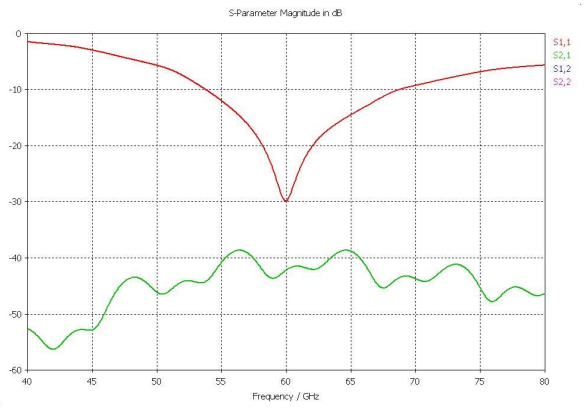
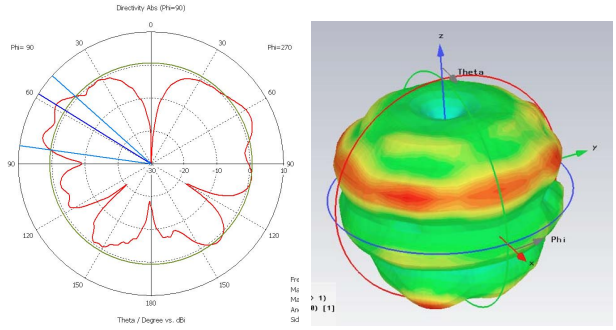


Fig. 10 Simulated S11 and S21 of the monopole antenna system.



(a) (b)

Fig. 11 (a) E-plane and (b) 3-D radiation patterns of the monopole antenna at 60 GHz.

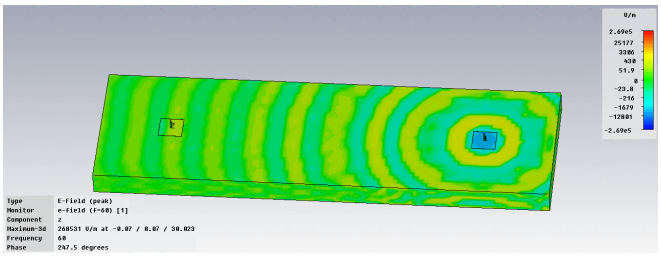


Fig. 12 Surface wave propagation of the monopole antenna system at 60 GHz.

Surface wave propagation: The surface wave of the antenna with a simplified skull plate at different angles of observation is demonstrated in Fig. 15 (a) to (c). From the sectioned side view in (a), we can observe the propagation of the wave inside the skull plate, although it is weaker than the propagation in the air. However, as the skin and brain, which are of much higher dielectric constant, have not been considered in the simulation, we believe the two high dielectric layers will confine the wave inside the skull and result in better propagation. More evidences will be given in the conference. Fig. 15 (b) and (c) show the top view of the wave, they show how the wave propagates to the desired direction (downward), this is the reason we propose to use more directional antennas in this application. The constructive and destructive interferences shown in (c) are caused by the strong reflections at the skull-air boundary. Modeling a bigger skull plate or including the complete skull model can be minimized these interferences.

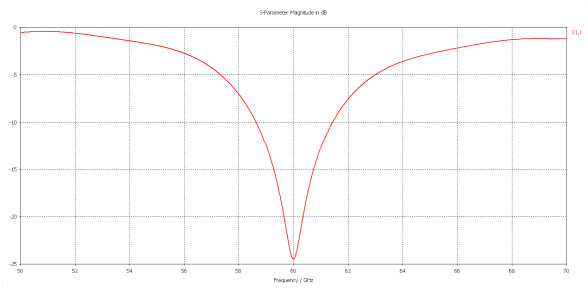
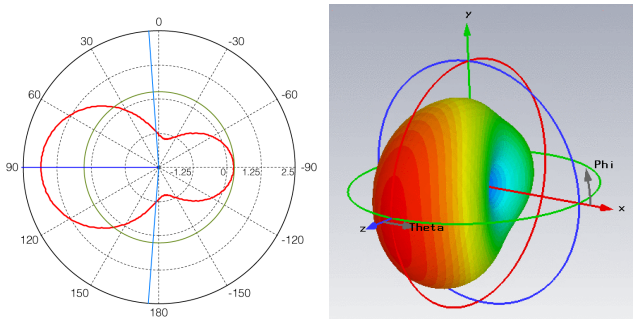


Fig.13 Simulated S11 of the SIW antenna.



(a) (b)

Fig. 14 (a) E-plane, and (b) 3-D radiation patterns of the SIW antenna at 60 GHz

B. Spike sorting hardware

The accuracy of Hebbian eigenfilter based spike sorting algorithm has been studied in our previous work [6]. This paper focuses on evaluating hardware results. Because of massive embedded computing resource, field programmable logic arrays (FPGAs) quite suits for implementing parallel hardware structure that is the case of multi-channel spike sorting. The reconfigurable capability makes FPGAs more flexible than specific hardware. In this paper, we use FPGAs to verify our hardware structure and explore different structure in design space. Xilinx Virtex6 XC6VLX760 is the target device. The simulation results of power and area consumption are obtained through Xilinx ISE design suite.

1) Real-time processing part

Fig. 16 (a) shows the relationship between folding ratio and number of consumed logic resources in terms of LUT (look-up table, the basic logic element in FPGA). In folded structures, area can be reduced by sharing computing resources among channels. However, storage units that cache intermediate results cannot be shared. As a result, as the number of folded channel increases, the area reduction will become less significant.

The dynamic power consumption of hardware with various folding ratio is shown in Fig. 16(b). In general, dynamic power is proportion to clock frequency and capacitance of circuits (area of circuit). For an N-fold multi-channel structure, the clock frequency should be N times faster than the fully parallel structure to enable real time processing. As folding ratio increases, the area reduction tends to saturate as shown in Fig. 16 (a). Therefore, power consumption per channel will mainly determined by the clock frequency and increase with the number of folded ratio.

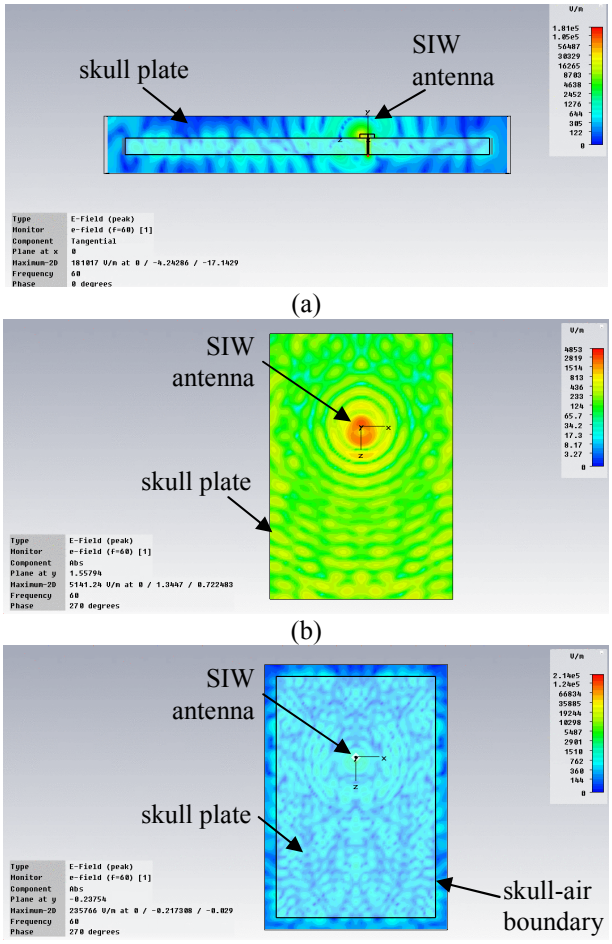


Fig. 15 Surface wave propagation of the SIW antenna system at 60 GHz (a) sessioned side view, (b) top view at the antenna level, and (c) top view inside the skull layer

The power consumption and logic resource cost have inverse tendency with increases of folding ratio. If considering power and logic resource together in terms of the production of power and logic resource, an optimal value of folding ratio is obtained, as shown in Fig. 15(c). In the optimal case, 8 channels are folded on a real-time processing core.

2) Training part

Hardware of Hebbian eigenfilter and k-means algorithm for parameter training is also designed. Table I lists the power and hardware resource consumption of the training hardware. We assume that multi-channel spike sorting hardware consists of one set of training hardware and multiple N-folded real-time processing units. The total hardware resource cost is formulated by equation (1) and (2), where N_{LUT}, N_{Train_LUT}, $N_{realtime\_LUT}$ and N_{ch} are the total number of consumed LUT, LUT consumption of training hardware, LUT consumption per channel of real-time processing hardware and the number of channel. P_{Train_LUT} and $P_{realtime\_LUT}$ are dynamic power consumption of training part and power per channel of real-time processing hardware.

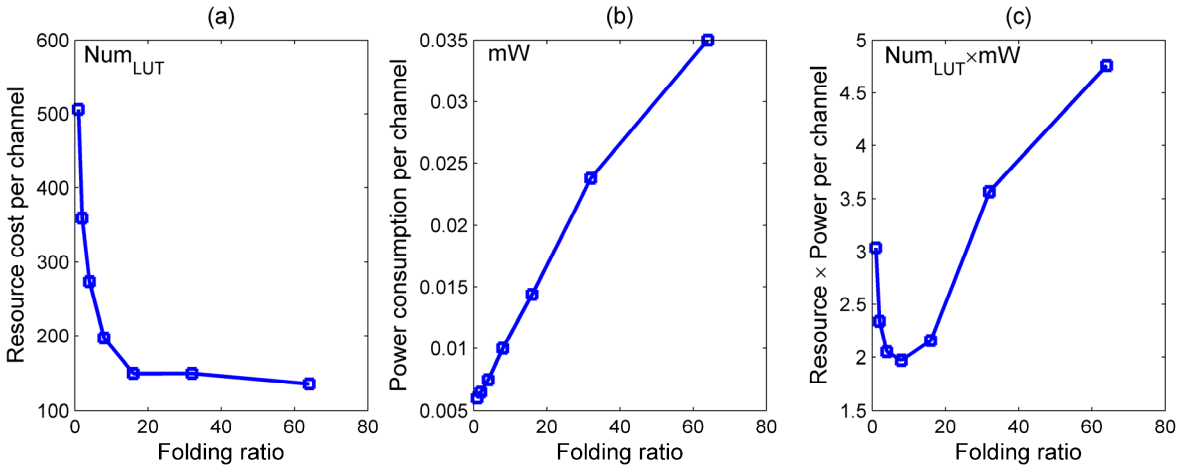


Fig. 16 Relationships between folding ratio and resource cost per channel (in terms of number of look-up table, LUT) (a), folding ratio and dynamic power consumption per channel (b) and folding ratio and production of resource and power consumption.

$$N_{LUT} = N_{Train\_LUT} + N_{ch} \times N_{realtime\_LUT} \quad (1)$$

$$Power = P_{Train\_LUT} + N_{ch} \times P_{realtime\_LUT} \quad (2)$$

Utilizing equation (1), (2) and power and logic resource consumption per channel, the total power and logic consumption of the FPGA can be obtained. Fig. 17 shows the power and logic resource consumption of an M channels spike sorting hardware.

Table I Power and hardware resource consumption

	Eigenfilter	K-means	Total
LUT	512	2172	2684
Power(mW)	5.4	7.2	12.6

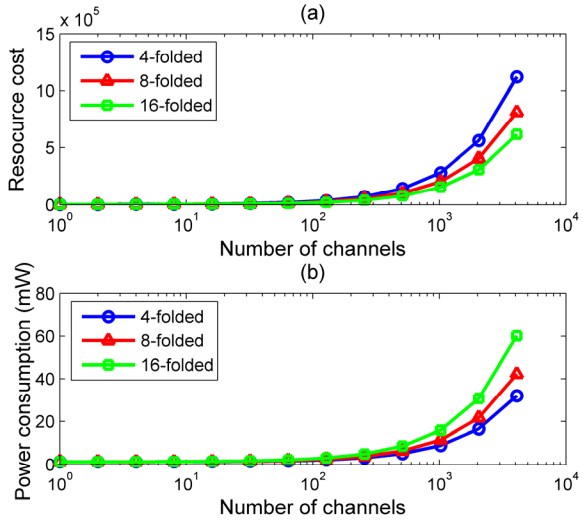


Fig. 17 Total resource cost (a) and power consumption (b) under various number of channels.

IV. Conclusion

This paper presents design explorations in antenna part and multi-channel spike sorting hardware of a recurrent BMI system. Three kinds of antennas are modeled. S-parameter, far field radiation and surface wave propagation of each antenna are studied. A Hebbian eigenfilter based spike sorting hardware is designed targeting on FPGA. Parallel and folding techniques are employed to make a trade of between area and power consumption. An optimal value considering both area and power consumption can be obtained through the empirical method.

References

[1] Jackson, A.; Mavoori, J.; & Fetz, E. E.; "Long-term motor cortex plasticity induced by an electronic neural implant". Nature 2006, 444(7115), 56-60.

[2] Kipke, D., Shain, W., Buzsáki, G., Fetz, E., Henderson, J. M., Hetke, J. F., and Schalk, G.; "Advanced neurotechnologies for chronic neural interfaces: new horizons and clinical opportunities", Journal of Neuroscience, 28 (11830), 11838, 2008.

[3] Tong K. F.; Luk K. M.; 'Design of microstrip mobile phone antennas', Digest 1994 Asia-Pacific Microwave Conference, Tokyo, Japan, December 1994. pp 1257-1259

[4] Bokhari, S.A.; Zurcher, J.-F.; Mosig, J.R.; Gardiol, F.E.; 'A small microstrip patch antenna with a convenient tuning option' IEEE Trans. on Antennas and Propagation, vol.: 44 , issue: 11, pages: 1521 - 1528, 1996

[5] CST Microwave Suite www.cst.com.

[6] Michael S Lewicki, "A review of methods for spike sorting: the detection and classification of neural action potetials," Network: Comput. Neural Syst. 9, 1998.

[7] Bo Yu, Terrrence Mak, Xiangyu Li, et al., "A Reconfigurable Hebbian Eigenfilter for Neurophysiological Spike Train Analysis," International Conference on Field Programmable Logic and Applications, 2010.

[8] Kyung Hwan Kim and Sung June Kim, "Neural spike sorting under nearly 0-db signal-to-noise ratio using nonlinear energy operator and artificial neural-network classifier," IEEE Transactions on Biomedical Engineering, vol. 47, no.10, 1406–1411, 2000.

[9] Simon S.Haykin, "Neural networks and learning machines," 3rd edition, Prentice Hall, 2009.

[10] Zach Zumsteg, Caleb Kemere, S. O'Driscoll, et al., "Power Feasability of Implantable Digital Spike-Sorting Circuits for Neural Prosthetic Systems," IEEE Trans. on Neural Systems and Rehabilitation, Vol. 13, No. 3, pp. 272-279, 2005.