INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Towards Two-point Neuron-inspired Energy-efficient Multi-modal Open Master Hearing Aid

M Raza[1], *A Adetomi*[1,2], *K Ahmed*[1], *A Hussain*[2], *T Arslan*[3], *A Adeel* [4]

[1]CMI Lab, University of Wolverhampton, Wolverhampton
[2]School of Computing, Edinburgh Napier University, Edinburgh
[3]School of Engineering, University of Edinburgh, Edinburgh
[4]School of Computing, Stirling University, Stirling

ahsan.adeel@deepci.org

Abstract

Here we demonstrate a two-point neuron-inspired audio-visual (AV) open Master Hearing Aid (OpenMHA) framework for on-chip energy-efficient speech enhancement (SE). The developed system is compared against state-of-the-art cepstrum-based audio-only (A-only) SE and conventional point-neuron inspired deep neural net (DNN) driven multimodal (MM) SE. Pilot experiments demonstrate that the proposed system outperforms audio-only SE in terms of speech quality and intelligibility and performs comparably to point neuron-inspired DNN with significantly reduced energy consumption at any time — both during training and inferencing.

Index Terms: OpenMHA, energy efficiency, two-point neurons

1. Introduction

Hearing aids (HAs) are the most widely used devices for compensating the majority of hearing losses. However, trouble understanding speech-in-noise (SIN) persists. There is a rapidly growing interest in developing new types of sustainable and reproducible assistive hearing systems. A few recent advancements include openMHA as a portable signal processing tool [1] which is widely used by the research community to analyse and improve commercial HA devices. Inspired by the human performance in everyday noisy situations, which is known to be dependent upon both aural and visual senses, here we introduce a novel two-point-inspired MM openMHA framework that uses visuals to effectively and efficiently clean noisy speech. Specifically, the contributions of this paper can be summarized as follows: (1) we present a two-point neuron-inspired openMHA framework for on-chip energy-efficient MM speech enhancement (2) We validate our approach using benchmark AV Grid [2] and ChiME3 [3] corpora, with 4 different real-world noise types (cafe, street junction, public transport (BUS), pedestrian area) and compare against conventional openMHA algorithms and conventional point neuron driven deep net. Comparative results demonstrate that our proposed method outperforms these methods in terms of energy consumption and SE performance.

2. Two-point neuron driven Open-MHA framework

A System-on-Chip (SoC) FPGA, which integrates both processor and programmable hardware fabric into a single device, provides a platform for targeting hardware-software co-processing in HA development. The openMHA framework, which is an open-source software platform for real-time audio signal processing, is set up on the processor, while the FPGA's hardware fabric is used to implement the video interface and audio interface subsystems. A layer of device drivers is added to allow the audio codec device to be exposed by the embedded Linux as a sound card for use by the openMHA. This hardware-software setup provides the base system for the implementation of a DNN-based streaming AV SE system.

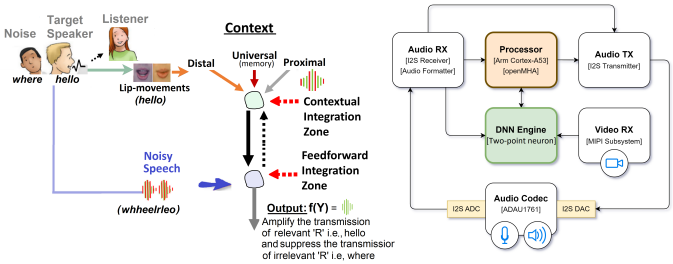

Figure 1: *Two-point neuron-inspired energy-efficient MM Open Master Hearing Aid on SoC FPGA: (left) functional depiction of two-point neuron [4][5]. (right) OpenMHA implementation.*

Audio Interface: The audio interface comprises of the I2S Receiver, Audio Formatter, and I2S Transmitter blocks, all implemented with Xilinx IP cores. These hardware blocks provide an interface for the I2S-based audio interface device on the SoC FPGA, with the I2S Receiver connecting to the ADC side while the I2S Transmitter connects to the DAC side. The audio interfacing is through the ADAU1761, which is a low-power stereo audio code with a sampling rate of up to 96 kHz. The device features a 24-bit ADC (Analog-to-Digital Converter) and DAC (Digital-to-Analog converter), along with flexible digital signal processing capabilities that include a 28-bit signal path and up to 50 MIPS of processing power. The I2S Receiver is the input stage that receives I2S audio data from the Audio Codec via a microphone. It supports I2S, left-justified, right-justified, and PCM (Pulse Code Modulation) audio formats. The audio Formatter is the processing stage that handles the incoming I2S data and performs digital signal processing (DSP) operations on it. It provides high-bandwidth direct memory access between the I2S Receiver and AXI4-Stream target peripherals supporting audio data, the I2S Transmitter in this case. The audio Formatter implements a variety of DSP functions, such as sample rate conversion, channel mixing, and filtering, allowing the customization of the processing pipeline. The I2S transmitter is the output stage that sends processed I2S data to the audio Codec, for onward routing to a speaker.

Video Interface: The Video RX is made up mainly of the Mobile Industry Processor Interface (MIPI) Camera Serial Interface (CSI-2) RX subsystem, which captures images from a Pcam 5C MIPI CSI-2 camera module and outputs AXI4-Stream video data ready for image processing. The Pcam

5C imaging module has onboard, the Omnivision OV5640 5-MP system-on-chip colour image sensor, and it transfers data over a dual-lane MIPI CSI-2 interface. It connects to the Soc FPGA board via a 15-pin flat-flexible cable, providing enough data bandwidth for video streaming at the following video formats: QSXGA@15Hz, 1080p@30Hz, 720p@60Hz, VGA@90Hz and QVGA@120Hz. The output of the camera is encoded by the Bayer pattern and in the RAW form, has the same format as a grey-level image. As such, there is no need for further processing as a grey-level image is sufficient for the lip-reading functionality. The video data is routed form the Video RX to the DNN Engine.

Embedded Linux Development: To properly set up the openMHA framework on the Arm core of the SoC FPGA, a Debian-based Linux Operating System (OS) is required, with the two main components of the Linux OS being the kernel, and the rootfs. For the Xilinx SoC boards, an easy way of developing embedded Linux OS is to use the PetaLinux. However, the PetaLinux rootfs is not compatible with the openMHA. It is, therefore, necessary to replace the PetaLinux rootfs with Debian-based rootfs. Moreover, to control the ADAU1761, a layer of device driver is added during the embedded Linux development. This allows the audio codec to be exposed as an ALSA sound card for use by the openMHA.

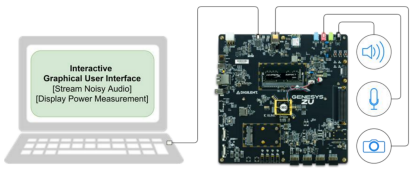

Figure 2: *Demonstration setup*

Two-point neuron engine for MM SE: The DNN engine exploits the deep context-sensitive neural information processing mechanism, termed multisensory cooperative computing (MCC), from [4][5] with 18 convolutional layers. The DNN engine uses both audio and visual cues for AV SE. The widely utilised Grid [2] and ChiME3 [3] corpora are the basis for our tests here. The output of the DNN engine feeds into the openMHA on the Processor via a plugin wrapper. This is an ongoing implementation.

Hardware Setup: The development is done on an SoC FPGA device from Xilinx. The Zybo AP SoC board has been used to implement an audio-only streaming openMHA as a precursor to the eventual DNN-based AV SE hearign aid demonstrator. The board shown in Fig. 3 is the Genesys ZU-3EG, which is a Zynq Ultrascale+ MPSoC development board with a xczu3eg-sfvc784-1-e chip.

3. Results

For energy estimations, the highly-distributed parallel implementation of our brain-inspired, non-von Neumann MCC architecture on a Xilinx UltraScale+ MPSoC device is used. In this architecture, a synaptic value of zero contributes nothing to the energy consumption. The dynamic power consumption of the MAC unit is 2 mW as reported by the XPower Estimator tool. The energy consumption due to an activated neuron is equivalent to 2 mW X 4 clock cycles X 10-ns period, which is equal to 0.08 nJ per synapse in a single feedforward run. This implies that an inactive neuron saves 0.08 nJ per single inference run. More details are comprehensively presented in [4][5].

The DNN models estimate the clean spectrogram when the noisy spectrogram and visual images are fed as input. The estimated magnitude is combined with the noisy phase to generate enhanced speech using an inverse STFT. Figure shown in demo video depicts the overall neural activity of the proposed context-sensitive two-point neuron driven MCC model and conventional DNN during training. It can be observed that context-sensitive processors in MCC quickly evolve to become highly sensitive to relevant information and activate only when the information is relevant for the task at hand. Therefore, significantly reduced neural activity is evident. Table 1 presents the performance comparison of MCC with the standard Cepstrum based A-only SE and point neurons-inspired deep net in terms of energy consumption, Perceptual Evaluation of Speech Quality (PESQ), and Short-Time Objective Intelligibility (STOI).

Table 1: *Inference: provisional estimated energy consumption per Grid utterance, PESQ, and STOI results. Audio-Only Cepstrum-OpenMHA vs Baseline-OpenMHA, and MCC-OpenMHA.*

SNR	Energy (J)			PESQ			STOI		
	Cepstrum	Baseline	MCC	Cepstrum	Baseline	MCC	Cepstrum	Baseline	MCC
-6dB	2.659	2652	**25.57**	1.03	1.68	1.64	0.34	0.62	0.63
0dB	2.659	2652	**25.57**	1.26	1.74	1.70	0.43	0.66	0.70
6dB	2.659	2652	**25.57**	1.53	1.79	1.77	0.48	0.75	0.77

4. Conclusion

The standard A-only OpenMHA is designed to be very low power with significant compromises on the speech quality and intelligibility, especially in extreme noisy environments. The deep net on the other hand requires improves speech quality and intelligibility but at the cost of high energy demands. This pilot study shows that the proposed two-point neuron-driven OpenMHA requires far less energy for MM SE compared to conventional point neuron-inspired DNN driven OpenMHA. We hypothesise that our proposed OpenMHA approach contributes towards on-the-fly online MM SE, targeting limited HA space and power budgets.

5. References

[1] C. Pavlovic, V. Hohmann, H. Kayser, L. Wong, T. Herzke, S. Prakash, z. Hou, and P. Maanen, "Open portable platform for hearing aid research," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1738–1738, 2018.

[2] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[3] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime'speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE ASRU*. IEEE, 2015, pp. 504–511.

[4] A. Adeel, M. Franco, and M. Raza, "Context-sensitive neocortical neurons transform the effectiveness and efficiency of neural information processing," *arXiv preprint arXiv:2207.07338*, 2022.

[5] A. Adeel, A. Adetomi, K. Ahmed, A. Hussain, T. Arslan, and W. Phillips, "Unlocking the potential of two-point cells for energy-efficient and resilient training of deep nets," *IEEE Transactions on Emerging Topics in Computational Intelligence*, pp. 1–11, 2023.