INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

5G-IoT Cloud based Demonstration of Real-Time Audio-Visual Speech Enhancement for Multimodal Hearing-aids

Ankit Gupta[1], *Abhijeet Bishnu*[2], *Mandar Gogate*[3], *Kia Dashtipour*[3], *Tughrul Arslan*[2], *Ahsan Adeel*[4], *Amir Hussain*[3], *Tharmalingam Ratnarajah*[2], *and Mathini Sellathurai*[1]

[1]Heriot-Watt Watt University, Edinburgh, United Kingdom
[2]University of Edinburgh, Edinburgh, United Kingdom
[3]Edinburgh Napier University, Edinburgh, United Kingdom
[4] University of Wolverhampton, Wolverhampton, United Kingdom.

{abishnu,t.ratnarajah, t.arslan}@ed.ac.uk, {ankit.gupta,m.sellathurai}@hwu.ac.uk,
{m.gogate, k.dashtipour, a.hussain}@napier.ac.uk, a.adeel@wlv.ac.uk

Abstract

Over twenty percent of the world's population suffers from some form of hearing loss, making it one of the most significant public health challenges. Current hearing aids commonly amplify noises while failing to improve speech comprehension in crowded social settings. In this demonstration, we showcase a proof-of-concept implementation of the world's first 5G and Internet of Things (IoT) enabled multi-modal hearing aid (MM HA) prototype. This integrates an innovative 5G cloud-radio access network (C-RAN) and IoT based transceiver model for real-time audio-visual speech enhancement (AVSE). Specifically, we demonstrate a transceiver model for Cloud-based AVSE which satisfies high data rate and low latency requirements for future MM HAs. The innovative 5G-IoT transceiver application is shown to satisfy HA latency limitations while transmitting raw noisy AV data from an MM HA prototype device to the cloud for deep learning-based real-time AVSE processing and obtaining a clean audio signal.

1. Introduction

Designing efficient, truly personalised hearing-aid (HA) technologies is a formidable interdisciplinary research and innovation challenge. Amongst other factors, the very low uptake of state-of-the-art HAs is attributed to performance limitations of multi-channel audio-only speech enhancement algorithms which are known to be ineffective in suppressing overwhelming background noise and improving speech comprehension in crowded social settings. Our recent works [1], [2] have demonstrated the potential of low-latency deep learning (DL) algorithms for designing multi-modal hearing-aid (MM-HA) devices that can contextually combine audio and lip movement video, to deliver significant speech intelligibility improvements over state-of-the-art audio-only hearing-aid (HA) devices. However, the requirement for low-powered and low-complexity HA devices makes the on-chip application of computationally-expensive DL-based audio-video (AV) speech enhancement (SE) algorithms infeasible. Thus, AV data from the MM-HA needs to be transmitted to the cloud for processing by AV SE algorithms. The latter may be implemented using smart glasses for real-time AV sensing and Cloud-based AVSE processing integrated with conventional audio-based HAs as sound delivery units. Thus, we design a novel 5G-IoT transceiver for MM-HA that integrates a unique 5G cloud-radio access network (C-RAN) and IoT based transceiver model for real-time AVSE to satisfy the challenging latency and high data rate requirements.

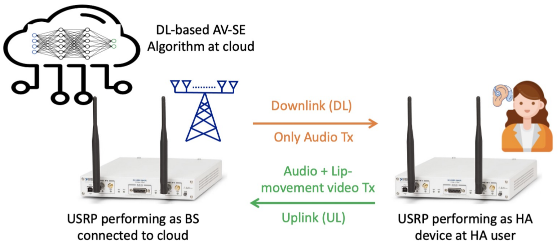

Figure 1: *Real-time cloud-based 5G-IoT framework for AV SE in MM-HA.*

Specifically, as shown in Fig. 1, we develop a novel off-chip transceiver prototype fora real-time cloud-based 5G-IoT framework for AV SE in MM-HA [3]. Herein, HA user's audio and lip-movement data is captured using a small camera and transmitted by the USRP to the cloud in real-time for processing by real-time AVSE algorithms [4]. Once processed, the cleaned audio is transmitted by another USRP in real-time to the HA user's USRP. For wireless channel transfer to the cloud, the data needs to be transformed into a frame structure.

In Sec. 2, we detail the design of the frame structure for the low-latency 5G-IoT transceiver. In Sec. 3, we detail the real-time AV SE algorithm utilized at the cloud. In Sec. 4, we analyze the end-to-end latency of the proposed transceiver. Finally, in Sec. 5, we conclude this demonstration with user experience.

2. Frame Structure Design of 5G-IoT Transceiver for Cloud-based MM-HA

We design a new frame structure for low-latency communication (based on our recent work [3]), detailed below.

The frame has a length of 10 ms and is made up of 10 subframes with a duration of 1 ms each. To reduce the latency in the transmission and reception of the signal by encoding and decoding of the frame structure, we propose to utilize frame structure with 60 OFDM symbols, i.e., 6 OFDM symbols in a frame. While, we keep the 15 kHz subcarrier spacing. In the uplink channel (from the HA user to the cloud) both audio and lip-movement video transmission is transmitted, thus we allow 1.4 MHz and 3 MHz bandwidth.

The downlink channel (from the cloud to the HA user), on the other hand, only transmits the enhanced audio signals, thus has a constant data rate and can accommodate 1.4 MHz of bandwidth. For 1.4 MHz (3 MHz) bandwidth, the 0th and 7th OFDM symbols in each subframe should have cyclic prefix (CP) lengths of 10 (20) and the remaining OFDM symbols should have CP lengths of 9 (18). Further, instead of using low density parity check (LDPC) codes we propose to utilize the Bhose-Choudary (BCH) codes for channel coding. This is because BCH codes are very powerful codes for shorter block lengths (as we have 6 OFDM symbols in a frame).

The downlink frame structure consists of PSS/SSS, PDCCH, PDCCH reference signal, DRS, and PDSCH. We also design the downlink control information of 25 bits, where we keep 10 bits for frame number, 1 bit for code-rate, 1 bit for modulation, 4 bits to denote number of frames in the transport block, 1 bit for end of payload, 3 bits for uplink SS ID and 5 bits are kept reserved.

There are four components to the uplink frame structure: PSS/SSS, PUCCH, PUCCH reference signal, and PUSCH. Please note URS and PUSCH are dependent on the selected bandwidth. We also design the uplink control information of 25 bits, where we keep 10 bits for frame number, 1 bit for code-rate, 1 bit for modulation, 4 bits to denote number of frames in the transport block, 1 bit for end of payload, 1 bit for bandwidth and 7 bits are kept reserved.

3. Audio-Visual Speech Enhancement Algorithms for MM-HA

The real-time AV SE method creates an optimal binary mask that reduces the noise dominant portions while enhancing the speech dominant portions from cropped lip pictures of the intended person and a noisy audio spectrogram. Our CochleaNet [4] based MM AV SE neural network design uses depthwise separable convolutions, lower STFT window size of 32 ms, narrower STFT window shift of 8 ms, and 64 convolutions in the speech feature extraction layers for lowering the computation latency. Additionally, a low-latency framework for visual extraction of features is used.

4. Real-time Latency Analysis

We perform transmission of AV data in uplink and enhanced audio data in the downlink (after processing using the AV SE algorithm detailed in the next section) and determine the latency of the designed 5G-IoT framework. We find that the latency[1] of frame structures in (1) downlink tranmission is 4.15 ms, (2) downlink reception is 56.92 ms, (3) uplink tranmission is 3.18 ms, and (4) uplink reception is 41.55 ms. Thus, end-to-end latency for the designed cloud-based 5G-IoT framework is 105.8 ms. Note that this latency is in software defined radio using USRP and ongoing hardware (FPGA) based implementation of the frame structure is leading to a 10 fold reduction. Thus, our innovative design frame structure is highly suitable for low-latency communication in the cloud-based 5G-IoT MM-HA system. Detailed latency measurements can be found in Table 1, where ”Each big frame” and ”Each small frame” denotes 140 and 6 OFDM symbols, respectively.

5. Conclusion and User Experience

We propose an innovative transceiver demonstration for cloud-based AV SE in MM-HAs that can handle streaming data frame-by-frame. As a result, Interspeech visitors will see the very first demonstration of a physical layer transceiver that delivers AV SE in real-time while adhering to rigorous HA latency and data rate constraints. We will bring a pair of computers and a pair of USRP x310 devices for this presentation.

Table 1: *Latency of different communication blocks in MM-HA.*

Blocks	**Time**	**Occurrence**
Transmitter blocks of DL frame structure		
SSS	18 μs	one
Cell RS Symbol	460 μs	Each small frame
Mapping	20 μs	Each small frame
PBCH Index	15 μs	Each big frame
PDSCH Index	40 μs	Each small frame
RE2BIN	325 μs	Each small frame
BCH Encoder	400 μs	Each small frame
Transport Block	450 μs	Each small frame
PBCH Symbol	320 μs	Each big frame
Data Generation	2.2 ms	Each small frame
Receiver blocks of DL frame structure		
SSS Det	22 ms	one
PBCH Decoder	8.7 ms	Each big frame
Channel Correction	365 μs	Each small frame
BCH Decoder	390 μs	Each small frame
Payload Detection	640 μs	Each small frame
Data Detection	24.5 ms	Each small frame
Transmitter blocks of UL frame structure		
SRS	20 μs	one
DMRS	19.5 μs	Each small frame
PBCH Index	3 μs	Each small frame
PUSCH Index	65 μs	Each small frame
PBCH Symbol	450 μs	Each big frame
Data Generation	2.6 ms	Each small frame
Receiver blocks of UL frame structure		
SRS Det	22 ms	one
PBCH Decoder	770 μs	Each big frame
Channel Correction	380 μs	Each small frame
CFO Correction	185 μs	Each small frame
Data Enhancement	15 ms	one
Data Detection & AV SE	2.2 ms	Each small frame

6. Acknowledgements

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) programme grant: COG-MHEAR (Grant reference EP/T021063/1). Website: `https://cogmhear.org/index.html`

7. References

[1] A. Adeel, J. Ahmad, H. Larijani, et al., “A Novel Real-Time, Lightweight Chaotic-Encryption Scheme for Next-Generation Audio-Visual Hearing Aids,” in *Springer Cognitive Computation*, vol. 12, pp. 589601, 2020, doi: 10.1007/s12559-019-09653-z.

[2] A. Adeel, A. Adetomi, K. Ahmed, A. Hussain, T. Arslan and W. A. Phillips, “Unlocking the Potential of Two-Point Cells for Energy-Efficient and Resilient Training of Deep Nets,” in *IEEE Transactions on Emerging Topics in Computational Intelligence*, pp. 1-13, 2023, doi: 10.1109/TETCI.2022.3228537.

[3] A. Bishnu, A. Gupta, M. Gogate, K. Dashtipour, A. Adeel, A. Hussain, M. Sellathurai and T. Ratnarajah, “A Novel Frame Structure for Cloud-Based Audio-Visual Speech Enhancement in Multimodal Hearing-aids,” in *IEEE International Conference on E-health Networking, Application & Services (HealthCom), Genoa, Italy*, pp. 75-80, 2022, doi: 10.1109/HealthCom54947.2022.9982772.

[4] M. Gogate, K. Dashtipour, A. Adeel, A. Hussain, “CochleaNet: A robust language-independent audio-visual model for real-time speech enhancement,” in *Information Fusion*, vol. 63, pp. 273-285, 2020, doi:10.1016/j.inffus.2020.04.001.

[1] The latency is measured on Intel (R) Core (TM) i7-9700 CPU @ 3.00 GHz processor, 64 GB RAM, and LabVIEW NXG 3.0.