Live Demonstration: Real-time Multi-modal Hearing Assistive Technology Prototype

Mandar Gogate, Adeel Hussain, Kia Dashtipour, Amir Hussain
School of Computing, Edinburgh Napier University, Edinburgh, UK
Email: {m.gogate,adeel.hussain,k.dashtipour,a.hussain}@napier.ac.uk

I. Background

Hearing loss affects at least 1.5 billion people globally. The WHO estimates 83% of people who could benefit from hearing aids do not use them. Barriers to HA uptake are multifaceted but include ineffectiveness of current HA technology in noisy environments with multiple competing noise sources where human performance is known to be dependent upon input from both the aural and visual senses.

In this live demonstration, we will showcase a first of its kind real-time, multi-modal speech enhancement prototype that can contextually exploit lip reading cues to effectively enhance speech in real noisy environments. This will serve to demonstrate the technical feasibility of developing real-time audio-visual (AV) deep neural network based algorithms for multi-modal hearing-assistive technology. Our transformative approach [1], [2], [3] will extract salient information from the pattern of the speaker's lip movements, whilst preserving speaker privacy, and contextually employ this information as an additional input to AV speech enhancement algorithms.

As part of the interactive hands-on demonstration, participants will communicate in real-time to assess the qualitative benefit of utilising novel AV speech enhancement models compared to conventional audio-only deep neural network based approaches. Participants will be able to select and experiment with a variety of real background noises and a range of AV models to learn how these can autonomously adapt to the nature and quality of visual and acoustic environmental inputs.

II. Demonstration Setup

An overview of the demonstration setup is depicted in Fig. 1. The setup consists of a web camera (to record target speaker), binaural headphones (to playback output) and noise sources (speakers and laptop will be used to generate multiple competing interference).

III. Visitor Experience

The visitors will witness AV speech enhancement in real-time through the live demonstration. Specifically, visitors will be able to experience and evaluate a significant improvement in the speech quality and intelligibility through our proposed algorithm's contextual use of video input and low-latency, privacy-preserving combination of audio and visual speech information.

An interactive graphical user interface will allow visitors to interact with the real-time prototype demonstration. The target speaker will be physically situated in a real noisy environment with a range of real-world (competing speech and non-speech type) environmental noises. The listeners will hear enhanced target speaker speech without any interfering noise in both online (MS Teams-based) and physical (one-to-one and group) communication scenarios.

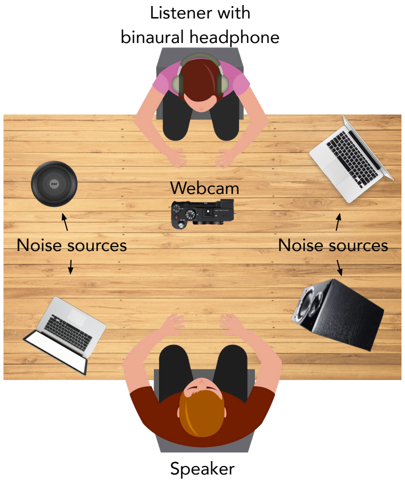

Fig. 1. Live Demonstration setup - AV hearing-assistive technology prototype

The live demonstration of our proposed real-time AV algorithm will be shown to overcome multiple challenges and constraints associated with state-of-the-art speech enhancement approaches, such as limited generalisation and lack of user privacy preservation. In addition to understanding the potential of utilising our approach in multi-modal hearing assistive technology, visitors will learn of benefits to exploit the prototype as a real-time integrated AV speech enhancement tool within their social and professional web-based meeting platforms.

IV. Acknowledgement

This work was funded by the UK Engineering and Physical Sciences Research Council (EPSRC) programme grant: COG-MHEAR (Grant Ref. No.EP/T021063/1)

References

[1] M. Gogate, K. Dashtipour, A. Adeel, and A. Hussain, “Cochleanet: A robust language-independent audio-visual model for real-time speech enhancement,” *Information Fusion*, vol. 63, pp. 273–285, 2020.

[2] M. Gogate, K. Dashtipour, and A. Hussain, “Visual speech in real noisy environments (vision): A novel benchmark dataset and deep learning-based baseline system.” in *Interspeech*, 2020, pp. 4521–4525.

[3] ——, “Towards real-time privacy-preserving audio-visual speech enhancement,” in *Proc. 2nd Symposium on Security and Privacy in Speech Communication*, 2022, pp. 7–10.