INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Application for Real-time Audio-Visual Speech Enhancement

*Mandar Gogate, Kia Dashtipour, Amir Hussain*

Edinburgh Napier University, Edinburgh, UK
{m.gogate,k.dashtipour,a.hussain}@napier.ac.uk

## Abstract

This short paper demonstrates a first of its kind audio-visual (AV) speech enhancement (SE) desktop application that isolates, in real-time, the voice of a target speaker from noisy audio input. The deep neural network model integrated in this application exploits the AV nature of speech from the target speaker to suppress all speech and non-speech background sounds. In the context of a growing need for video conferencing solutions, AV SE enables the practical deployment such technology in challenging acoustic environments with multiple competing background noise sources. In these scenarios, classical audio-only SE typically fails as they are usually trained to isolate speech from non-speech noises. The application comprises a graphical user interface and modules for real-time AV speech acquisition, preprocessing, and enhancement. The participants will experience a significant improvement in the speech quality and intelligibility of a target speaker who will be physically situated in a real noisy environment with a range of real-world noises. Moreover, participants can evaluate the performance of the application with their own voice by recording videos in challenging multi-talker conversational environments.
**Index Terms**: speech enhancement, audio-visual speech separation, real-time processing, multimodal communications

## 1. Introduction

The COVID-19 outbreak has had a significant impact on communication and collaboration practices. Video conferencing applications have been widely adopted for personal and business communication due to the need of social distancing and remote work [1]. The quality of such communication is adversely affected by background noise, reverberations, and interfering speakers. This significantly increases the cognitive load and reduces the speech intelligibility of the listener [2]. In such situations, state-of-the-art audio-only speech enhancement (SE) algorithms are commonly exploited to suppress the non-speech background noise. However, in challenging multi-talker environments, the interfering speaker cannot be effectively suppressed by using audio-only SE.

In order to address the aforementioned issue, audio-visual (AV) SE models have been proposed to isolate target speaker's voice from a mixture of speech and non-speech noises [3]. AV SE models exploit the lip movements of the target speaker to suppress unwanted background interference as the lip motion is tightly correlated with the acoustic information. Alternatively, personalised SE can be used if a sample of clean speech from the target speaker is available. This paper focuses on exploiting AV SE models to isolate target speakers speech for real-time video communications.

Specifically, we develop an application based on our real-time AV SE models [4, 5] that utilises lip information of the target speaker and noisy spectogram. The application generates a spectral mask that enhances target speech dominant regions and suppresses noise dominant regions in real time. The model is trained to isolate users voice from interfering speech, music and other non-speech noise sources.

The proposed application can be used to preprocess audio for any AV communication app, such as Zoom, Skype, Microsoft Teams, FaceTime and WhatsApp, and can also be utilised for low-latency multimodal hearing assistive technologies. It enables better communication than state-of-the-art audio-only SE methods in challenging acoustic environments including busy cafeterias, trains and restaurants.

## 2. Algorithms

### 2.1. Model architecture

In the literature, extensive research has been conducted to develop AV SE models [3]. However, limited work has been carried out to develop real-time AV SE models. In this application, we exploit our previously proposed real-time AV SE model including privacy preserving AV SE model [5] and CochleaNet [4] for real-time video communication. The visual feature extraction of CochleaNet is replaced with ShuffleNet v2 0.5x [6]. In addition, the window size and shift for STFT is also reduced to 400 and 128 respectively. It is to be noted that, both models are causal and can be used for processing streaming data (frame by frame).

### 2.2. Data preprocessing

The audio signal is downsampled to 16kHz and segmented into 32 ms frames with 8 ms increment. Short-time Fourier transform (STFT) and hanning window is applied to generate 257 bin spectrum. For visual data, the preprocessing pipeline presented in [5] and CochleaNet [4] was used without modification.

### 2.3. Model training

The models were trained using the benchmark AVSE Challenge dataset [7]. The dataset is augmented with noises from Audioset [8] and LibriSpeech [9]. The clean speech is mixed with randomly selected noise at SNR ranging from -15 dB to 5 dB. Subjective listening tests using real noisy VISION corpus [10] are conducted to evaluate the model performance in speaker and noise independent settings.

## 3. Desktop application

The desktop application takes streaming audio from microphone, video from webcam and process it using the AV SE

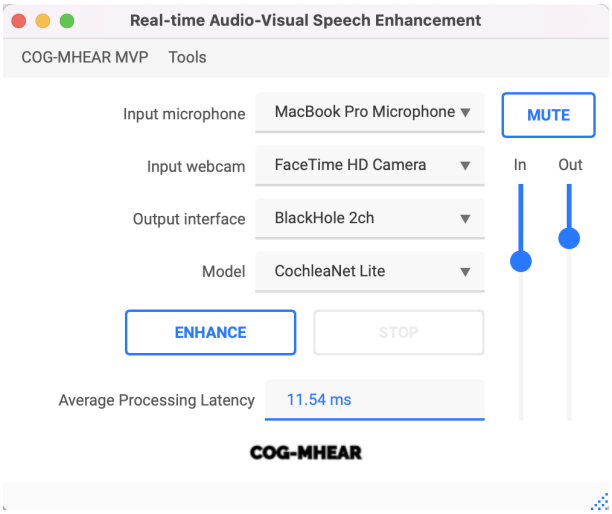


Figure 1: *Application graphical user interface*

model. The enhanced audio can be routed using any video conferencing application using a virtual microphone such as Blackhole[1]. The application is implemented using the Qt framework[2] and works on Mac, Linux and Windows. The model inference is performed using ONNX runtime library[3]. The audio pipeline includes getting the data from the microphone, downsampling, STFT, model inference, ISTFT, upsampling and writing the audio to a loopback device. The visual pipeline includes reading from the webcam, face detection, landmark point and lip extraction. It is to be noted that, a single frame of AV data is fed to the SE model at a time. Since, the audio is sampled at a higher sampling rate than visuals, visual frames are upsampled to match the audio sampling rate.

The application user interface is shown in Fig. 1. Users need to select the interface for AV input and virtual loopback interface for writing the processed data. On the right, input and output levels of interfaces can be set to the appropriate value. The application also includes a record and playback feature to hear the difference between noisy and enhanced speech. The total latency on a M1 Pro Macbook is approximately 12 ms plus small loopback interface latency. This does not affect the video conferencing communication quality.

## 4. Conclusions

In this paper, we presented a desktop application that isolates users voice in the presence of multiple competing speech and non-speech noises. The application uses a real-time audio-visual (AV) speech enhancement (SE) model that exploits lip movements to enhance users voice and suppress all other background noise sources. The model addresses the limitation of audio-only SE methods to distinguish between target and distractor speech. The proposed application can be used with any video conferencing application (e.g. Zoom, Microsoft Teams, Google Meet) to suppress annoying noise sources when calling from busy social environments. It can also be exploited for low-latency multimodal hearing assistive technologies.

The presented desktop application is a research prototype that is still under development as part of COG-MHEAR project[4]. The application has several limitations that will be addressed in the future. First, the AV dataset used for training only contains English. To make the application more widely applicable, more languages should be added to the AV dataset. Second, the application's CPU usage and energy footprint are relatively high. To reduce these, the AV SE model should be optimized for inference by quantizing the model and/or pruning it. These challenges provide a roadmap for future work on the application.

[1]https://github.com/ExistentialAudio/BlackHole
[2]https://www.qt.io
[3]https://onnxruntime.ai

## 5. Acknowledgements

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

## 6. References

[1] R. K. Sandhu, J. Vasconcelos-Gomes, M. A. Thomas, and T. Oliveira, "Unfolding the popularity of video conferencing apps–a privacy calculus perspective," *International Journal of Information Management*, vol. 68, p. 102569, 2023.

[2] Z. Zhu, H. Yang, M. Tang, Z. Yang, S. E. Eskimez, and H. Wang, "Real-time audio-visual end-to-end speech enhancement," *arXiv preprint arXiv:2303.07005*, 2023.

[3] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, 2021.

[4] 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.

[5] M. Gogate, K. Dashtipour, and A. Hussain, "Towards real-time privacy-preserving audio-visual speech enhancement," in *Proc. 2nd Symposium on Security and Privacy in Speech Communication*, 2022, pp. 7–10.

[6] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[7] A. L. A. Blanco, C. Valentini-Botinhao, O. Klejch, M. Gogate, K. Dashtipour, A. Hussain, and P. Bell, "Avse challenge: Audio-visual speech enhancement challenge," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 465–471.

[8] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[9] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[10] 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.

[4]https://cogmhear.org