



Evaluating the Audio-Visual Speech Enhancement Challenge (AVSEC) Baseline Model Using an Out-of-Domain Free-Flowing Corpus

Kia Dashtipour¹, Bryony Buck¹, Mandar Gogate¹, Adeel Hussain¹, Arif Reza Anwary¹, Tughrul Arslan², Amir Hussain¹

¹ School of Computing, Engineering and the Built Environment, Edinburgh Napier University, UK

² School of Engineering, The University of Edinburgh, Edinburgh, UK

k.dashtipour@napier.ac.uk

Abstract

The human auditory cortex contextually integrates audio-visual (AV) cues to enhance the comprehension of speech in noisy environments. Numerous studies have investigated the effectiveness of AV integration for speech enhancement (SE). This paper evaluates the effectiveness of the COG-MHEAR AV SE Challenge baseline model using an out-of-domain free-flowing corpus. Experimental results indicate that the COG-MHEAR AV SE Challenge baseline model exhibits superior performance when applied to an out-of-domain corpus.

Index Terms: Free-Flowing, Speech enhancement, Audio-Visual

1. Introduction

The primary objective of speech enhancement (SE) is to enhance speech intelligibility and quality in noisy signals by attenuating noise components. Over recent decades, numerous SE methods have been proposed and demonstrated to significantly improve sound fidelity. One prominent technique, spectral restoration, involves estimating a gain function to suppress noise components in the frequency domain and thereby reconstruct a cleaner speech spectrum from noisy input [1, 2]. More recently, deep learning models have been increasingly employed in SE to achieve superior performance in noise reduction tasks. Deep learning-based SE methods generally surpass conventional approaches in effectiveness. Approaches utilizing recurrent neural networks and convolutional neural networks have particularly demonstrated promising results in this domain.

Gogate et al. [3] propose a novel language, noise, and speaker-independent audio-visual (AV) deep neural network architecture for real-time SE. The model utilizes noisy acoustic cues and robust visual cues to enhance speech intelligibility by selectively focusing on the target speaker. Their approach exhibits superior performance, as demonstrated on datasets such as the GRID corpus and CHiME 3 noise conditions. In addition, Gogate et al. [4] presented a deep neural network (DNN) based on an AV mask estimation model. This model contextually integrates the temporal dynamics of both audio and visual features for mask estimation and speech separation. The AV feature extraction and ideal binary mask estimation utilize a hybrid DNN architecture that leverages the complementary strengths of stacked long short-term memory (LSTM) and convolutional LSTM networks. Initial simulation results demonstrate the superior performance of this approach.

In this paper, we evaluate AVSEC baseline model using an out-of-domain free-flowing corpus [5]. The free-flowing corpus exhibits a more natural and less scripted quality compared to the AVSEC data on which the model was initially trained.

This new dataset contains unscripted, free-flowing, AV conversational data between three interlocutors of mixed hearing ability. In evaluating the AVSEC model using data with high levels of realism, we aim to assess the model's limitations when exposed to naturalistic AV speech.

2. Methodology

This section describes the COG MHEAR AVSE challenge baseline model.

2.1. Audio Feature Extraction

Audio features were extracted utilising a U-net network, comprising modified encoder and decoder blocks specifically designed for AVSE. The input to the network is the magnitude of the noisy speech short-time Fourier transform (STFT) spectrogram, characterised by frequency and time dimensions. This input is processed through two convolutional layers with a filter size of 4 and a stride of 2, facilitating down-sampling of the time-frequency dimensions until the time dimension is reduced to 64. Subsequently, the down-sampled features are passed through three convolutional blocks, each consisting of two convolutional layers with a filter size of 3 and a stride of 1, followed by a frequency pooling layer that reduces the frequency dimension by a factor of two.

2.2. Visual Feature Extraction

The visual feature extraction component of the pipeline initiates with a 3D convolutional layer, employing a filter size of $5 \times 7 \times 7$ and a stride of $1 \times 2 \times 2$, followed by the ResNet-18 architecture [6]. The output features from the residual network are subsequently processed by a temporal convolutional network (TCN). The input to the network comprises a sequence of lip-cropped images with dimensions $N \times 88 \times 88$, where N represents the number of frames. For each lip image, the visual feature network generates a 512-dimensional vector. These visual features are then upsampled to correspond with the audio feature sampling rate.

2.3. Multimodal fusion

The upsampled visual and audio features are integrated and fed into a U-Net decoder. The decoder consists of three up-convolutional blocks, each comprising two upsampling layers that double the time dimension, followed by convolutional layers with a filter size of 3 and a stride of 1. Subsequently, these AV features are processed through two transposed convolutional layers with a filter size of 4 and a stride of 2, progressively expanding the time-frequency dimension to match the original input size. A sigmoid layer is then employed to map the output to

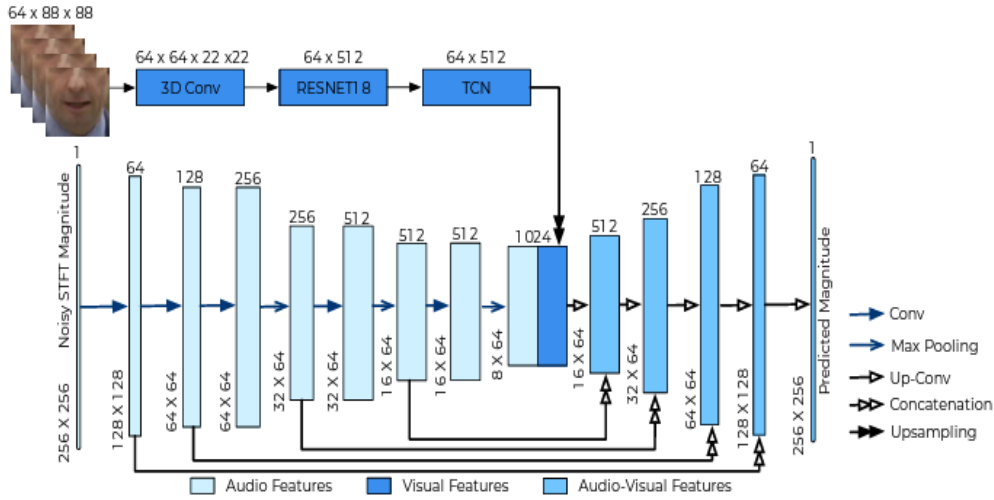


Figure 1: The FCN-based U-Net framework used to optimize the STOI-based audio-visual SE model

a range between 0 and 1. Finally, the predicted mask is applied to the input spectrogram to produce the masked spectrogram as the output. Figure 1 depicts the FCN-based U-Net framework used for the COG-MHEAR AV SE baseline model.

3. Experimental Results and Discussions

In this section, we provide an overview of the free-flowing corpus, the objective evaluation metrics, and the experimental setup, followed by a discussion of the results.

3.1. Free Flowing Corpus

Participants with mixed hearing abilities were recruited from the Edinburgh area and invited to participate in groups of three for short (2-3 minute) prompted conversations. Each group consisted of one bilateral hearing aid user and two non-hearing aid users. Conversations were conducted in both quiet and noisy environments in a sound attenuated room. For the purpose of this investigation, conversations held in a multilayered soundscape comprising multi-talker babble, music, and environmental sounds at 76 dBA were used. Conversation prompts were designed to evoke discursive conversation, aiming to facilitate inclusive turn-taking and to avoid topic-related monologue or withdrawal (extending [7, 8]). The corpus data collection included audio recordings from lavalier lapel microphones (Zoom F2) worn by each participant, and 2D video recordings of each participant’s face and shoulders, as well as video captured from the hearing aid user’s perspective (video recorded using ELP USB Webcams at 1920x1080 dpi and 30 fps, collated using OBS recording software). Audio and visual recordings were synchronized and segmented using Audacity and OpenShot video editing software (version 3.1.1). Corpus design and data collection are presented in greater detail in [5]. This



Figure 2: Example visual outputs from free-flowing conversation corpus.

dataset presents novel multi-talker speech data recorded with high levels of realism, not currently available in existing AV speech datasets.

3.2. Objective testing on synthetic mixtures

The evaluation of speech processing quality often involves subjective listening tests. In our experimental framework, we utilize various evaluation metrics, namely PESQ, STOI, and SI-SDR, to approximate subjective evaluations. PESQ computes a linear combination of average disturbance values and asymmetrical disturbance values between a reference signal and its modified counterpart. However, PESQ primarily assesses one-way speech distortion and noise speech quality, lacking representation of interactive effects such as loudness, loss, delay, sidetone, and echo. Scores for PESQ range from [-0.50, 4.50], indicating the potential range of reconstructed speech quality [9]. STOI

quantifies the correlation of short-time temporal envelopes between clean and modified speech, with values ranging from [0, 1], where higher values denote enhanced intelligibility [10]. Lastly, SI-SDR, a scale-invariant variant of SDR, measures the distortion introduced by the separated signal relative to clean signal energy. Higher SI-SDR values reflect superior speech separation performance [11].

3.3. Experimental Results

We investigated the impact of using an out-of-domain free-flowing corpus on the AV SE challenge baseline model in terms of PESQ, STOI, and SI-SDR. Table 1 summarises the results. After model application, the PESQ is 2.41 compared to 1.12 prior to AVSE. The STOI is 0.83 following AVSE, compared to 0.61 before enhancement. Finally, the SI-SDR rose to 8.06 from -0.47 before model application. These experimental findings demonstrate that the free-flowing out-of-domain corpus achieved better performance following AVSE challenge model application.

Table 1: Results for out-of-domain Free-Flowing corpus

	PESQ	STOI	SI-SDR
Noisy	1.12	0.61	-0.47
Free-Flowing out-of-domain	2.414	0.83	8.06

4. Discussion

Previously, the AVSE challenge model has been trained and tested using raw and pre-processed datasets generated using AV single utterances extracted from TED and TEDx videos [12]. Testing with out-of-domain data allows evaluation of how well the model generalises to previously unseen conditions, not covered in the training scenarios. In this instance, the AVSE challenge model was evaluated using novel recordings of free-flowing multi-talker conversations held in complex noise environments.

Applying the AVSE method to the out-of-domain data resulted in improved scores for measures of speech quality (PESQ), intelligibility (STOI), and noise distortion (SI-SDR). The increase in PESQ score following AVSE application, shows a notable improvement in perceptual speech quality. Furthermore, the observed rise in STOI score suggests a substantial improvement in speech intelligibility due to the AVSE model. Additionally, we observed a substantial rise in SI-SDR value, demonstrating a notable reduction in signal distortion after applying the AVSE model to the free-flowing corpus data. As such, the AVSE model can be seen to generalise well to unseen AV speech data with high levels of realism.

These findings highlight the effectiveness of incorporating visual elements of conversation when enhancing speech signals, particularly in challenging environments where traditional audio-only methods might falter. This is consistent with findings in the field that highlight the benefits of multimodal approaches to speech processing [13][14].

5. Conclusion

In this paper, we assess and evaluate the COG-MHEAR AV SE baseline model using an out-of-domain free-flowing corpus. The initial experimental results demonstrate that the baseline model achieves promising results. As ongoing future work, we

intend to evaluate the model’s performance using different languages and conversation settings to understand its generalisation capabilities.

6. Acknowledgments

The authors are grateful to the anonymous reviewers for their invaluable comments and suggestions. The authors acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC) Grants Ref. EP/T021063/1 (COG-MHEAR) and EP/T024917/1 (NATGEN); The Royal Society of Edinburgh (2977); and The Royal Society (IES \R2 \222037)

7. References

- [1] J. Chen, J. Benesty, Y. Huang, and E. J. Diethorn, “Fundamentals of noise reduction,” *Springer Handbook of Speech Processing*, pp. 843–872, 2008.
- [2] P. Scalart *et al.*, “Speech enhancement based on a priori signal to noise estimation,” in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2. IEEE, 1996, pp. 629–632.
- [3] 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.
- [4] M. Gogate, A. Adeel, R. Marxer, J. Barker, and A. Hussain, “Dnn driven speaker independent audio-visual mask estimation for speech separation,” *arXiv preprint arXiv:1808.00060*, 2018.
- [5] B. Buck, K. Dashtipour, M. Gogate, A. Q. Hussain, A. L. A. Blanco, M. A. Akeroyd, and A. Hussain, “Advancing multimodal hearing-aid processing: An audio-visual corpus of free-flowing conversations,” in *International Hearing Aid Research Conference, August 20-25 California, USA. IHCON, 2024*.
- [6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [7] B. Buck, A. McLaren, and G. Naylor, “Worse hearers make better listeners: The effects of hearing loss on turn-taking and head movement in virtual conversations,” in *6th International Conference on Cognitive Hearing Science for Communication, June 12-15, Linköping, Sweden. CHSCOM, 2022*.
- [8] K. D. A. H. Bryony Buck, Mandar Gogate, “Freeflow: A github repository,” 2024, accessed: 2024-08-09. [Online]. Available: <https://github.com/kiadashtipour/freeflow.git>
- [9] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.
- [10] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Transactions on audio, speech, and language processing*, vol. 19, no. 7, pp. 2125–2136, 2011.
- [11] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “Sdr-half-baked or well done?” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.
- [12] 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.
- [13] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Deep audio-visual speech recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 8717–8727, 2018.

- [14] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multi-modal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.