AVSE Challenge: Audio-Visual Speech Enhancement Challenge

AVSE CHALLENGE: AUDIO-VISUAL SPEECH ENHANCEMENT CHALLENGE

Andrea Lorena Aldana Blanco[1], *Cassia Valentini-Botinhao*[1], *Ondrej Klejch*[1], *Mandar Gogate*[2], *Kia Dashtipour*[2], *Amir Hussain*[2], *Peter Bell*[1]

[1]University of Edinburgh, Edinburgh, UK
[2]Edinburgh Napier University, Edinburgh, UK

ABSTRACT

Audio-visual speech enhancement is the task of improving the quality of a speech signal when video of the speaker is available. It opens-up the opportunity of improving speech intelligibility in adverse listening scenarios that are currently too challenging for audio-only speech enhancement models. The Audio-Visual Speech Enhancement (AVSE) challenge aims to set the first benchmark in this area. We provide participants with datasets and scripts to test their audio-visual speech enhancement models under a common framework for both training and evaluation. The data is derived from real-world videos, and comprises noisy mixes, in which audio from target speaker is mixed with either a competing speaker or a noise signal. The submitted systems are evaluated by conducting AV intelligibility tests involving human participants. We expect this challenge to be a platform for advancing the field of audio-visual speech-enhancement and to provide further insight about the scope and limitations of current AV speech enhancement approaches.

Index Terms— Audio-visual speech enhancement, subjective intelligibility, LRS3 dataset

1. INTRODUCTION

Poor speech intelligibility is a major barrier to effective human communication. Intelligibility can be easily degraded by environmental factors affecting the signal such as background noise and the presence of competing speakers; this may particularly affect hearing-impaired listeners. Speech enhancement aims to overcome such degradation by improving a target speech signal in terms of intelligibility and quality. In most cases, speech enhancement models focus on suppressing background noise; however, in other cases, they focus on optimising the enhanced speech signal according to listener-specific characteristics. Speech enhancement can be used as a pre-processing step for downstream tools such as automatic speech recognition, or as an end-user application, for example, in hearing aids or public address systems.

To date, most speech enhancement models are audio-only (AO) [1, 2], meaning that they merely consider the audio signal as input. Furthermore, even though great progress has been made in recent years in the field of speech enhancement, every-day scenarios such as the case of a competing speaker mixed with a target speaker still represent a great challenge for AO models [3], particularly in cases in which both speakers are the same gender [4]. However, in a real-world scenario, it is normal for people to have access to visual information too – and previous studies have shown that visual cues such as lip-reading contribute to intelligibility [5]. This motivates the development of audio-visual (AV) speech enhancement models that are able to consider both audio and visual signals as input [6, 7, 8]. AV models have the potential to use context from the visual modality – particularly related to lip movement – that may be particularly helpful in selectively enhancing only the target speaker.

In this first edition of the Audio-Visual Speech Enhancement (AVSE) Challenge we address the problem of enhancing speech signals in two proposed scenarios that are challenging to AO models: (i) target speaker mixed with competing speaker and (ii) target speaker mixed with noise. In this edition we consider normal hearing conditions, but use signal-to-noise ratios (SNRs) that would conceivably be considered too challenging in AO conditions.

Previous work on AO speech enhancement evaluation has emphasized the importance of considering subjective evaluation of quality and intelligibility [9], and has compared how results from listening tests correlate to objective metrics, finding that existing metrics are not necessarily good predictors [10]. Based on these findings, audio-only speech enhancement challenges such as the Clarity Challenge [11], the Deep Noise Suppression (DNS) Challenge [12], and the Hurricane Challenge [13, 14] (for near-end listening enhancement) adopted listening tests as part of their main evaluation protocol. Additionally, there was a previous effort in conducting an audio-visual speech enhancement challenge [15]. However, the project developed into an evaluation of an audio-visual speech enhancement model and four AO models. System's performance was evaluated in terms of quality. To the best of our knowledge, no subjective evaluation protocols that account for speech intelligibility have been specifically designed for audio-visual speech enhancement. This is particularly surprising considering that existing objective metrics make their predictions based on audio only and

978-1-6654-7189-3/22/$31.00

therefore ignore the visual component.

Inspired by this we propose to evaluate (and therefore rank) challenge submissions based on subjective AV intelligibility tests. The design of the evaluation allows us to run listening tests both with and without participants having access to the video modality. We expect to not only provide better insight about protocols for AV speech enhancement evaluation, but also to work towards developing AV objective metrics to predict quality and intelligibility.

2. DATASETS

2.1. The challenge scenario

Participants address the task of enhancing a target speech signal that is mixed with an interferer. For the target signal we provide both the audio and video of the target speaker. The audio signal of the target speaker is mixed with an interferer that can be either (i) a competing speaker or (ii) noise. Target and interferer signals were mixed following the frequency weighted SNR calculation adopted in the Clarity Challenge [11]. In mixes including an interfering speaker, the target speaker is not explicitly identified to participants, but only the target speaker is included in the video. Systems must process each sample independently, so it is not possible to use information about either speaker from any other samples. In a future edition of the challenge we might consider introducing multi-speaker videos. It is important to note that even though we are not simulating reverberation in this challenge, both the target and the competing speech are derived from TED talks and therefore contain some level of reverberation.

2.2. Target speakers

The videos of the target speakers are selected from the LRS3 dataset [16]. LRS3 contains thousands of spoken sentences/phrases from TED and TEDx videos[1] of public lectures, each lecture being around ten minutes in length, and generally delivered by a single speaker. Sentences are segmented based on punctuation marks (commas, full-stops and question marks). The dataset provides cropped faces of speakers with a resolution of 224x224 pixels and a frame rate of 25 frames per second (fps). To construct the target speech for the training set we selected all speakers that had at least nine minutes of data. To construct the development set we randomly selected the remaining speakers that had at least five minutes of data. The audio track derived from the videos is monophonic and sampled at 16 kHz and 16 bits.

Unlike other speech enhancement challenges, where target material is derived from read speech datasets, our target material is more expressive (and potentially "clearer") as it was produced to be consumed by a live audience.

[1] https://www.ted.com/

2.3. Interferers

Audio tracks of interferers are composed of a single competing speaker or a noise source. All files are single channel with a 16 kHz sampling frequency and 16 bits of bit depth. Audio that originally had a higher sampling rate were downsampled to meet the above mentioned criteria.

2.3.1. Competing speaker

Competing speakers were randomly selected from the LRS3 dataset (for training and development set) and from more recent TED talks (for the evaluation set). Speakers were selected excluding the talks chosen as target such that talks from target and competing speakers are a disjoint set. To create the competing speaker recordings, audio extracted from all videos of each competing speaker was concatenated.

2.3.2. Noise

To create the noise dataset we collected audio files from three different sources:

- Clarity Challenge (First edition) [11]: common domestic noises derived mainly from Freesound[2] divided into 7 categories. We selected all files from all categories.
- DEMAND [17]: multi-channel recordings of soundscapes. We selected one channel of the following soundscapes: NFIELD, NPARK, NRIVER, OHALLWAY, OOFFICE, PCAFETER, PRESTO, PSTATION, SCAFE, SPSQUARE, STRAFFIC, TBUS, TCAR, TMETRO. We did not use the soundscapes labeled as DKITCHEN, DLIVING and DWASHING to avoid overlapping with domestics sounds from the Clarity Challenge, and soundscape OMEETING because of its resemblance to a competing speaker scenario.
- DNS Challenge (Second edition) [12]: sounds from AudioSet,[3] DEMAND and Freesound. We selected the Freesound noises (all 7 categories excluding "Fan" to avoid overlapping with the Clarity Challenge dataset).

The files in our noise dataset are divided into 15 categories: Dishwasher, Fan, Kettle, Hairdryer, Microwave, Vacuum, Washing machine, Soundscape, Breath, Copy-machine, Door, Dragging, Munching, Squeak and Typing.

2.4. Signal-to-noise ratio

To select appropriate SNR ranges for the training and dev sets we carried out a pilot evaluation. This included video mixes of the hardest (stationary noise) and easiest (competing speaker of a different gender) listening conditions at various

[2] https://freesound.org/

[3] https://research.google.com/audioset/

SNRs. Based on the results of this test we chose the following ranges: -15 dB to 5 dB (competing speaker) and -10 dB to 10 dB (noise). We note that these are particularly challenging conditions compared to those typically used in audio only evaluations [11], as we find that participants perform much better in audio visual intelligibility tests than in audio-only tests.

To avoid ceiling and flooring effects during subjective evaluation we selected three SNR values out of each range to create the evaluation set. This was done according to the procedure adopted in the Hurricane Challenge [13] that consisted on the estimation of psychometric curves based on word accuracy scores derived from a listening test. Figure 1 shows the curves obtained for each interferer derived from an online listening test involving 40 native English speakers. Participants were asked to type what they heard after watching each video clip. Videos contained 7-10 words sentences. Word accuracy scores were calculated following [13]. Based on these curves we chose SNR values of -9.3, -1.2 and 6.9 dB (noise) and -13.5,-5.4 and 2.7 dB (competing speaker) that reflect word accuracy scores of 25%, 50% and 75% (noise) and 40%, 55%, and 70% (competing speaker).

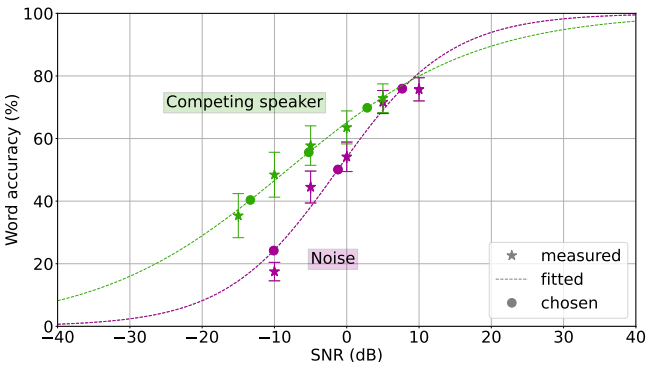

Fig. 1. Psychometric curves fitted on measured human word accuracy scores for competing speaker (green) and noise (purple). Error bars depict 95% confidence intervals.

2.5. Training and development sets

Training and development datasets are disjoint regarding target and competing speakers as well as noise files (the same noise categories are however present in both datasets). An overview is provided in Table 1.

Each mix has a unique sentence that is mixed with an interferer signal (i.e., noise, competing speaker) at a specific SNR. The selection of interferer type is uniform so that there is a similar number of mixes in the competing speaker scenario as there are noise mixes. Competing speakers and noise categories are randomly selected. For a chosen noise category, noise files are randomly selected considering that they need to be at least the same duration as the target speaker sentence. Given a competing speaker or noise file a segment of recording is randomly selected. Finally the SNR level is randomly sampled from the pre-defined interferer's SNR range.

	# Mixes	# Target speakers	Interferers
Train	34,524	605	405 competing speakers and 7,346 noise files.
Dev	3,306	85	30 competing speakers and 1,825 noise files.

Table 1. Training and development sets

2.6. Evaluation set

To create the evaluation set we selected a set of TED and TEDx talks that were not part of LRS3. The motivation for selecting new videos was to evaluate human intelligibility performance using the full-videos of the target speaker (not only the cropped faces) which are not available from LRS3. Moreover, we wanted to use unseen material (not yet released to the public) as part of the evaluation.

After selecting the videos, we extracted sentences based on the manual transcriptions of the talks. Afterwards, we processed the data using a modified version of the lip-synchronisation pipeline [18] to extract a set of sentences after we:

- discarded sentences that were shorter than two seconds;
- skipped sentences that contained more than one shot – we aimed for sentences in which the speaker face was visible and the shot was preserved throughout;
- discarded sentences that contained more than one face-track – we wanted to remove sentences with multiple speakers and sentences for which the face is not visible all the time;
- eliminated sentences for which it was not possible to lip-sync more than 80% of their duration. We used this parameter in the segmentation process to avoid issues such as the mouth of the speaker not being visible in the shot or the face-track not corresponding to the speaker.

The resulting evaluation set contains 1,389 extracted sentences from 30 speakers (15 females and 15 males). We provide face-tracks with a resolution of 224x224 and a frame rate of 25 fps. The dataset is balanced so that approximately half of the mixes have a competing speaker scenario and the other half noise. Competing speakers are selected from a pool of 6 competing speakers (3 females and 3 males).

2.6.1. Evaluation set: noises

Noises from the evaluation set belong to four categories that are a subset of the ones used in the train and dev sets. To

select the noise categories, we asked a native speaker to listen to a set of mixes including all noise categories and to rate the noises based on how difficult it was to understand the target speaker. Noises were classified according to three labels: easy, medium, hard. Then, we selected one noise assigned to each of the previous labels (easy, medium, hard). From the noises labeled as 'hard', we selected two examples: a stationary noise and a non-stationary noise. The noise categories included in the eval dataset are: microwave (easy), washing-machine (medium), hairdryer (hard), soundscape (hard).

Recordings were taken from Freesound. To avoid overlapping with sounds in the training and development sets, we only selected sounds that were uploaded after those used in the noise sources mentioned in section 2.3.2.[4] Moreover, we recorded additional noise samples from categories in which there were few recordings available in Freesound. The motivation behind the selected noise categories was to evaluate a wide range of noises among those used in train and dev, while taking into consideration time constraints derived from conducting listening test (i.e., noise fatigue). Table 2 shows an overview of the noise data collected for the eval set.

Category	# Recordings (mm:ss)	Source
Microwave	8 (05:57)	Freesound/ own recordings
Washing-machine	6 (10:07)	Freesound
Hairdryer	5 (04:25)	Freesound/ own recordings
Soundscape	11 (07:27)	Freesound

Table 2. Noise collection used in evaluation dataset

2.7. AVSE Challenge material

We provide participants with scripts for generating the training and development sets. Even though we are ranking the systems based on subjective AV intelligibility tests, we also provide evaluation scripts so that participants test performance of their systems using common objective intelligibility and quality metrics such as the short-time objective intelligibility (STOI) [19] and the Perceptual Evaluation of Speech Quality (PESQ) [20].[5] Both are intrusive metrics, the computation of which requires access to the clean signal. STOI is a metric designed to predict intelligibility from noisy speech. It is computed by estimating the linear correlation coefficient from the time-frequency representation of the clean and normalized noisy signals across time frames. This representation is obtained via one-third octave frequency band analysis. STOI values range between zero and one. PESQ is a metric initially designed to predict speech quality across telecommunication networks. As a first step the speech and noisy signals are equalized to match standard listening levels and then band-passed filtered according to the frequency response of the telecommunication channel. Then, the absolute difference in loudness spectra for both signals is calculated to then predict a value between 0.5 and 4.5 that resembles that of the Mean Opinion Score (MOS) scale.

[4]We gathered noises that were uploaded to Freesound after September of 2021.

[5]`https://github.com/cogmhear/avse_challenge`

3. EVALUATION PROTOCOL

Participant's submissions were evaluated according to word accuracy scores that were obtained from subjective evaluation involving human participants. We recruited 95 participants through the Prolific Academic platform from a pool of native British speakers with no self reported hearing difficulties and normal (or corrected to normal) vision. Participants are asked to perform the test in a quiet environment, wear headphones and to use either a desktop machine or a laptop. Validation videos are randomly placed throughout the test and the performance on these is used to exclude participants data during result analysis (data from 8 participants were excluded).

Following a similar protocol adopted in the Hurricane Challenge [13], participants were presented with mixes ordered in terms of interferer type and SNR such that the difficulty of the test progressively increases with time. Each video could only be played once and participants must input their responses in order to move forward.

Unlike conventional intelligibility evaluation that uses spoken content specially designed for intelligibility measurement (such as the Harvard sentences [13] or the matrix sentences [14]), our evaluation relies on pre-existing material (the TED talks). Because we have no control over how word confusability varies in a sentence rather than asking participants to type all words they can hear (and score each word, or keyword, equally) we adopt a different strategy. For each word in the evaluation set we find a set of similar words using phonetic similarity distance. In each utterance we replace a word whose worst alternative achieves the best perplexity computed with a 3-gram language model. For the evaluation we pick 120 utterances with the lowest 4-gram recall and we ask the participants to select among five alternatives. One of the alternatives is "none of the above", which is the correct alternative 20% of time. Such protocol minimises participant's listening effort enabling the evaluation of naturally occurring pre-existing material. This is particularly important for AV evaluation due to the scarcity of purposely designed AV data that can serve both for training SE models and for human evaluation. Prior to the challenge evaluation we validated this protocol against a transcription task on the same material. A full description of this and further details on the evaluation protocol can be found in [21].

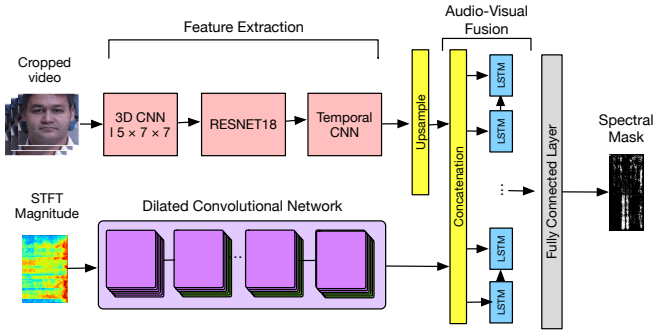

Fig. 2. Baseline Model

	conv1	conv2	conv3	conv4	conv5
Num filters	64	64	64	64	64
Filter size	5 x 5	5 x 5	5 x 5	5 x 5	1 x 1
Dilation	1 x 1	2 x 2	4 x 4	8 x 8	1 x 1

Table 3. Audio Feature Extraction

4. BASELINE SYSTEM

4.1. Data Preprocessing

The audio signals are sampled at 16 kHz and mono-channel is used for processing. The audio signal was segmented into 32 ms frames with 8 ms frame increment. The hanning window and Short-time Fourier transform (STFT) were applied to generate a 257-bin magnitude spectrogram. The videos are sampled at 25 frames per second and were used without any preprocessing as model input.

4.2. Model Architecture

The audio feature extraction module contains dilated convolutional layers as detailed in Table 3. Each of the layers is followed by a non-linear ReLU activation function. The dilated convolutions aggregates the multi-scale contextual information for the dense prediction problem. In addition, dilated convolutions are used for exponential expansion of receptive field without loss of coverage or resolution. The audio feature network outputs a 1028-D vector for each STFT frame.

The visual feature extraction stage of the pipeline consists of 3D convolutional layers with filter size of $5 \times 7 \times 7$ and stride of $1 \times 2 \times 2$, followed by ResNet-18 [22]. The residual network features are fed into temporal convolutional network. The input to the network is a time-series of face cropped images of size $N \times 224 \times 224$, where N is the number of frames. The visual feature network output a 512-D vector for each face image.

The visual features ($N \times 512$) are upsampled to match the audio feature sampling rate (T). The audio ($T \times 1028$) and upsampled visual features ($T \times 512$) are fused across time dimension to generate features of dimension $T \times 1540$. The fused features are fed into a LSTM layer which consists of 257 units. The LSTM output ($T \times 257$) is then fed into fully connected layers with 257 neuron and sigmoid activation. The weights of the fully connected layer are shared across the time dimension. The output of fully connected layer is multiplied with noisy speech features to generate the masked magnitude. The mean absolute error between masked magnitude and clean magnitude IBM is used as a loss function for network training. The complete pipeline is shown in Figure 2.

4.3. Experimental Setup

The baseline model is developed using Pytorch and a NVIDIA RTX A6000 GPU with 48 GB memory was used to train the model. The model is trained to minimise the mean absolute error with Adam optimiser (lr=16e-3) for 25 epochs. The learning rate is multiplied by 0.8 when the model validation loss stops decreasing for 2 consecutive epochs. The model with best validation loss is used for evaluation.

5. TECHNICAL SYSTEMS AND RESULTS

5.1. Entries

We included 10 entries in the evaluation, including **original** non-enhanced and the **baseline**. The other 8 were systems submitted to the challenge. One of them, the **AVSE01**, is a PyTorch implementation of the baseline model (written originally PyTorch lighting), that was trained across multiple machines using PyTorch's Distributed Data Parallel.

The **CogBiD** entry is based on U-Net [23] (an encoder-decoder convolutional model). U-Net's encoder consumes magnitude spectra derived from noisy speech. The encoded representation is then concatenated with image features processed by convolutional subnets including ResNet-18, and upsampled. The concatenated representation is fed to the decoder that estimates a multiplicative mask. Enhanced waveform is reconstructed from a masked noisy magnitude spectra and the noisy phase. The model is trained to maximise a modified version of STOI. **SLT AVSE** entry follows a similar architecture, with slight changes in number of layers and feature size, and the addition pose-invariant lip landmark flow features to the encoded video representation. These set of additional features are also used by the **ENU AVSE** entry, that instead of U-Net, uses a time-domain encoder-decoder architecture based on cross-attention. Additionally a series of transformer modules are used to process the concatenated audio and video representations prior to the system's decoder that estimates the enhanced waveform directly. To create **ENU AVSE 2** contrastive entry the output is further processed by Audacity's telephone equalization filter to remove artefacts.

The remaining entry **BioASP_CITI** is based on the deep complex convolution recurrent network [24] extended for AV enhancement. Noisy's speech complex spectra is processed

Entry	Name	Overall	Speech	Noise
A	Original	59.00	61.88	56.13
B	Baseline	52.30	54.41	50.19
C	AVSE01	50.29	57.09	43.49
D	SLT_AVSE	50.19	51.53	48.85
E	ENU_AVSE	66.57	83.72	49.43
F	ENU_AVSE_2	68.77	80.65	56.90
G	BioASP_CITI	66.19	79.12	53.26
H	BioASP_CITI_CE1	65.23	71.84	58.62
I	BioASP_CITI_CE2	63.31	72.22	54.41
J	CogBiD	52.68	52.87	52.49
LSD	3.35	4.55	4.73	

Table 4. Word accuracy scores (%) calculated across all conditions (Overall) and per masker (Speech and Noise).

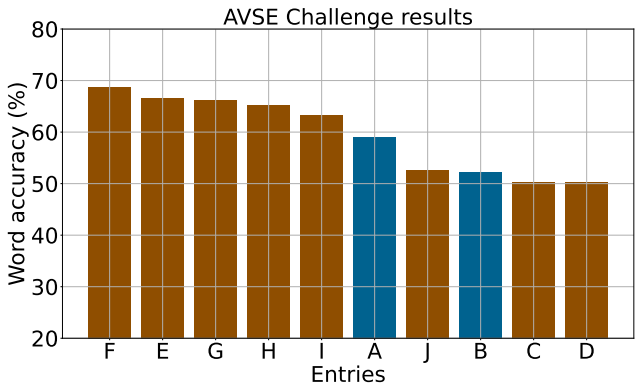

Fig. 3. Word accuracy (%) calculated across all maskers (LSD=3.35%). Original (A) and baseline (B) are in blue.

by a complex encoder and combined with ResNet-18 encoded video features via multihead cross-attention. The combined representation is processed by a complex decoder that predicts a complex multiplicative mask. The ISTFT operation is used to reconstruct the waveform from the masked complex spectra of the noisy signal. To create the contrastive entries **BioASP_CITI_CE1** and **BioASP_CITI_CE2**, noise augmentation is adopted during training (within mini batch). The probability to conduct noise augmentation increases with epoch up to 0.6 (CE1) or is fixed to 0.5 (CE2).

5.2. Results and discussions

Results from the AV intelligibility study are presented in Table 4. A higher score means more intelligible. Scores were calculated as the percentage of words correctly identified, computed for each participant and averaged across participants. Differences larger than the Fisher's least significant difference (LSD) are significant.

Fig. 3 presents the sorted overall results and Fig. 4 shows the significant differences between systems as solid boxes. The highest scoring submissions (F, E, G) are significantly different than the original but not significantly different from each other. Submissions J, C and D are significantly worse than the original (A). Results per masker presented in Table 4 show that all entries performed much better in the competing speaker masker than in noise. This is a surprising result given that audio-only speech enhancement tends to perform poorly in a competing speaker scenario. The highest scoring entry in noise (H) was the only entry that reported noise augmentation during training. The poor results in noise might reflect the fact that none of the entries were able to generalise to the evaluation set. Listening test participants might also have benefited more from the visual modality in the competing speaker case. This can only be confirmed with a further audio-only evaluation.

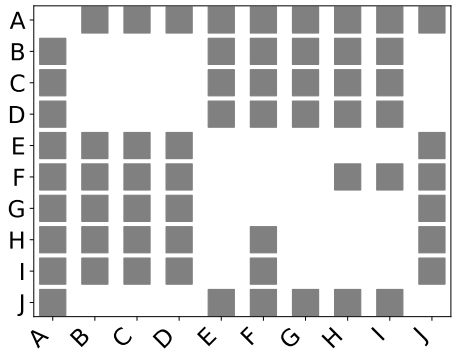

Fig. 4. Significant differences between the overall results each entry obtained are indicated by solid boxes.

6. CONCLUSIONS

We provided participants with datasets (train, development and evaluation) to test their speech enhancement models under a common framework. We evaluated the submitted systems based on AV intelligibility tests. Results indicate that there is room for improvement of audio-visual speech enhancement models. Further user studies are needed to explore if there are additional benefits of the visual input in the competing speaker scenario. Evaluation results will be presented as part of the Speech and Language Technology (SLT) workshop in January 2023.

Acknowledgements This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) programme grant: COG-MHEAR (EP/T021063/1).

7. REFERENCES

[1] 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 Trans. on Audio, Speech and Lang. Proc.*, vol. 29, pp. 1368–1396, 2021.

[2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," 2017.

[3] N. Saleem and M. I. Khattak, "A review of supervised learning algorithms for single channel speech enhancement," *Int. Journal of Speech Tech.*, vol. 22, pp. 1051–1075, 2017.

[4] Y. Z. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *CoRR*, vol. abs/1607.02173, 2016.

[5] S. Puschmann, M. Daeglau, M. Stropahl, B. Mirkovic, S. Rosemann, C. Thiel, and S. Debener, "Hearing-impaired listeners show increased audiovisual benefit when listening to speech in noise," *NeuroImage*, 2019.

[6] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, 2018.

[7] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *CoRR*, vol. abs/1804.04121, 2018.

[8] W. Wang, C. Xing, D. Wang, X. Chen, and F. Sun, "A robust audio-visual speech enhancement model," in *Proc. ICASSP*, 2020.

[9] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. ICSLP*, 1998, pp. 2819–2822.

[10] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 16, no. 1, pp. 229–238, 2008.

[11] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. V. Muñoz, "Clarity-2021 Challenges: Machine Learning Challenges for Advancing Hearing Aid Processing," in *Proc. Interspeech*, 2021, pp. 686–690.

[12] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "ICASSP 2021 Deep noise suppression challenge," in *Proc. ICASSP*, 2021.

[13] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Comm.*, vol. 55, no. 4, pp. 572–585, 2013.

[14] J. Rennies, H. Schepker, C. Valentini-Botinhao, and M. Cooke, "Intelligibility-enhancing speech modifications - The Hurricane Challenge 2.0," in *Proc. Interspeech*, 2020.

[15] M. Gogate, A. Adeel, K. Dashtipour, P. Derleth, and A. Hussain, "AV Speech Enhancement Challenge using a Real Noisy Corpus," 2019.

[16] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," in *arXiv preprint arXiv:1809.00496*, 2018.

[17] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. ICA*, 2013.

[18] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 251–263.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010.

[20] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001.

[21] A. L. A. Blanco, C. Valentini-Botinhao, O. Klejch, and P. Bell, "Efficient intelligibility evaluation using keyword spotting: a study on audio-visual speech enhancement," in *submission*, 2022.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *In Proc. MICCAI*, 2015.

[24] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Interspeech*, 2020.