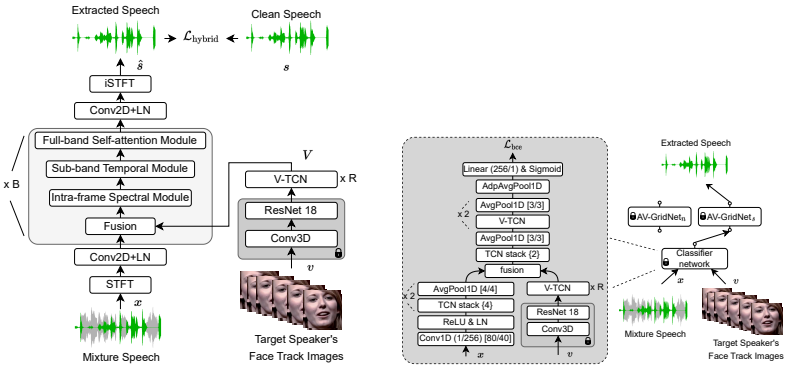
MERL's submission to the 2nd COG-MHEAR Audio-Visual Speech Enhancement Challenge

We participated in tracks 1 and 2 of the 2nd COG-MHEAR Audio-Visual Speech Enhancement Challenge. For each track, we use the result from SAV-GridNet as the primary submission, and the results from AV-GridNet as the contrastive submission. This report presents the system descriptions of our AV-GridNet and SAV-GridNet.

Network: Built upon the state-of-the-art speech separation model TF-GridNet [1], we propose AV-GridNet, a visually-grounded variant that incorporates the face recording v of a target speaker as a conditioning factor during the extraction process, shown in Fig. (a). We also build a scenario-aware version of AV-GridNet, named SAV-GridNet and shown in Fig. (b), that uses a classifier network to identify whether the mixture speech input is speech+speech or speech+noise, and then applies a dedicated expert model trained specifically for that scenario.



(a) AV-GridNet model (b) SAV-GridNet model

Table 1: Evaluation set results for track 1

Model	PESQ	STOI	SI-SDR [dB]
Noisy	1.14	0.44	−5.1
Baseline	1.41	0.56	3.7
AV-DPRNN[4]	1.94	0.73	10.4
AV-GridNet	2.59	0.83	13.8
SAV-GridNet	2.71	0.84	14.5
+ post-proc	2.71	0.84	14.5

(c) Benchmark results

The AV-GridNet hyper-parameter settings follow [1], except that the Conv2D, ResNet 18, V-TCN, and fusion layers follow [2]. The SAV-GridNet has an additional classifier network, in which the parameters follow the SLSyn network [3], except that the Conv2D, ResNet 18, V-TCN, and fusion layers follow [2]. The number of parameters is 9.8 M for AV-GridNet, 11.2 M for the combination of ResNet 18 and Conv3D, and 4.8 M for the classifier.

Training: We train AV-GridNet end-to-end with a hybrid loss [4], which consists of an SI-SDR loss and a frequency-domain delta spectrum loss. SAV-GridNet is a cascaded model consisting of a classifier network and two expert models (AV-GridNet$_n$ and AV-GridNet$_s$) that are trained independently. The classifier network is trained with binary cross-entropy loss. For all model training, we use the Adam optimizer with an initial learning rate of 0.001, the learning rate is halved if the best development loss (BDL) does not improve for 6 consecutive epochs, and the training stops when the BDL does not improve for 20 consecutive epochs. Dynamic mixing is also used. We train the model on 8 GPUs with 48 GB of RAM each. To fit the data in the GPU memory during training, the audio clips are truncated to 3 seconds for AV-GridNet, and 25 seconds for the classifier network. AV-GridNet has an effective batch size of 32, and is trained with 30 epochs that roughly take 4000 seconds each. The classifier network in SAV-GridNet has an effective batch size of 16, and is trained with 50 epochs that roughly take 600 seconds each. For models trained for track 2, we used all the speech utterances from the LRS3 dataset.

SAV-GridNet post-processing: For SAV-GridNet, we find that a model trained only on speech interference generalizes well on noise interference, but not vice versa. Hence, if the classifier predicts noise while the ground-truth label is speech and the wrong expert model is used, the results may be detrimental. Therefore, if the classifier prediction is noise and the criterion ($\mathcal{L}_{\text{SI-SDR}}(\hat{s},\hat{s}_n) < \mathcal{L}_{\text{SI-SDR}}(\hat{s},\hat{s}_s)$ or $\mathcal{L}_{\text{SI-SDR}}(x,\hat{s}_n) > \mathcal{L}_{\text{SI-SDR}}(x,\hat{s}_s)$) is met, which indicates that either the universal model is more in agreement with the noise expert than with the speech expert, or the original mixture is further to the output of the noise expert than to that of the speech expert, we keep the classification of the interference as noise, otherwise we change it to speech, where $\hat{s}_n = \texttt{AV-GridNet}_n(x,v)$, $\hat{s}_s = \texttt{AV-GridNet}_s(x,v)$, $\hat{s} = \texttt{AV-GridNet}(x,v)$, and $\mathcal{L}_{\text{SI-SDR}}(s,\hat{s}) = -20\log_{10}(\|\frac{<\hat{s},s>}{\|s\|^2}s\|/\|\hat{s}-\frac{<\hat{s},s>}{\|s\|^2}s\|)$.

Reference:

[1] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Making time-frequency domain models great again for monaural speaker separation," in *Proc. ICASSP*, 2023.

[2] Z. Pan, M. Ge, and H. Li, "USEV: Universal speaker extraction with visual cue," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 3032–3045, 2022.

[3] Z. Pan, R. Tao, C. Xu, and H. Li, "Selective listening by synchronizing speech with lips," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1650–1664, 2022.

[4] Z. Pan, M. Ge, and H. Li, "A hybrid continuity loss to reduce over-suppression for time-domain target speaker extraction," in *Proc. Interspeech*, 2022.