DA-AVSE: AN AUDIO-VISUAL SPEECH ENHANCEMENT SYSTEM WITH MULTI-MODEL MIXTURE PSEUDO-LABEL DOMAIN ADAPTATION METHOD

Chenxi Wang[1], *Hang Chen*[1], *Qing Wang*[1], *Jun Du*[1],
Chenyue Zhang[1], Xiaofei Ding[2], Feijun Jiang[2], Su Yue[2], Chin-Hui Lee[3],

[1]University of Science and Technology of China, Hefei, China
[2]Alibaba Group, Hangzhou, China [3]Georgia Institute of Technology, Atlanta, USA

1. PROPOSED SYSTEM

We use four models in the challenge: the official baseline model, MEASE [1], MTMEASE [2] and PLMEASE, which is a model that applies the audio-visual progressive learning framework proposed in [3] to MEASE. We propose a simple and efficient domain adaptation method called Multi-Model Mixture Pseudo-Label Domain Adaptation (MMMP-DA). The method is model-agnostic and can be applied to multiple models. The method jointly uses pseudo-labels generated by multiple models to fine-tune the models. The structure of the MMMP-DA method is shown in Figure 1.

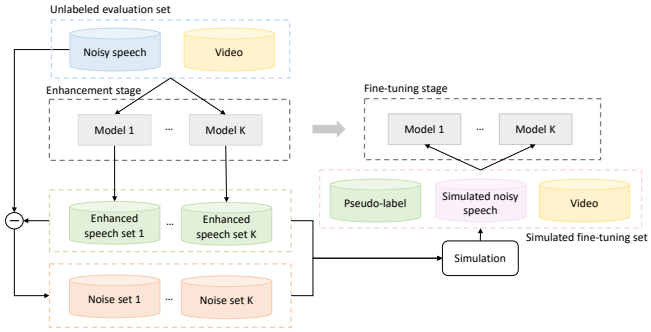

Fig. 1: The proposed MMMP-DA method.

We apply the MMMP-DA method to the above four models trained on the official training set. Specifically, the four trained models are used to predict enhanced speech of evaluation set. We regard the enhanced speech as pseudo-labels. Next, we subtract the enhanced speech from the noisy speech to obtain the noise. Each model will get a set of enhanced speech and noise predicted for the evaluation set. We put the enhanced speech of all models in one set and the noise in the other set. Using the enhanced speech set and the noise set, we simulate a 20-hour dataset for finetuning. When simulating each sample, we randomly select an enhanced speech and a noise, and the SNR is randomly selected from the range of 10 $\sim$ 0dB. Finally, we finetune the four models using this simulated dataset. We use the 10-fold cross-validation method during finetuning.

Furthermore, we use a multi-model fusion strategy to fuse the enhanced speech predicted by four models, resulting in more robust and reliable predictions.

2. EXPERIMENTS RESULTS

Table 1 presents the results of the official baseline, MEASE, MTMEASE and PLMEASE models without and with MMMP-DA method on the evaluation set. It can be seen that the MEASE, MTMEASE and PLMEASE demonstrated competitive performance across all evaluation metrics. After applying the MMMP-DA method, the performance of the four models all increased significantly, which demonstrate that our MMMP-DA method can effectively leverage the complementarity among multiple models and enhance their adaptation ability to the evaluation set.

Table 1: Performance comparison of official baseline, MEASE, MTMEASE and PLMEASE without and with the MMMP-DA method on the evaluation set.

Model	w/o MMMP-DA			w MMMP-DA		
	PESQ	STOI(%)	SISDR	PESQ	STOI(%)	SISDR
Baseline	1.41	55.63	3.67	1.63	65.51	8.43
MEASE	1.60	67.66	5.34	1.68	70.30	6.26
MTMEASE	1.61	67.86	5.46	1.68	70.30	6.24
PLMEASE	1.56	66.28	4.97	1.64	69.25	5.93

We fused the enhanced speech of the four models that applied the MMMP-DA method on the evaluation set to obtain the final result of our DA-AVSE system, as shown in Table 2.

Table 2: Performance of official baseline, DA-AVSE on the evaluation set.

System	PESQ	STOI(%)	SISDR
Noisy	1.14	44.10	-5.07
Baseline	1.41	55.63	3.67
DA-AVSE	**1.77**	**71.23**	**7.68**

3. REFERENCES

[1] Hang Chen, Jun Du, Yu Hu, Li-Rong Dai, Bao-Cai Yin, and Chin-Hui Lee, "Correlating subword articulation

with lip shapes for embedding aware audio-visual speech enhancement," *Neural Networks*, vol. 143, pp. 171–182, 2021.

[2] Chenxi Wang, Hang Chen, Jun Du, Baocai Yin, and Jia Pan, "Multi-task joint learning for embedding aware audio-visual speech enhancement," in *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2022, pp. 255–259.

[3] Chen-Yue Zhang, Hang Chen, Jun Du, Bao-Cai Yin, Jia Pan, and Chin-Hui Lee, "Incorporating visual information reconstruction into progressive learning for optimizing audio-visual speech enhancement," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.