A Study on Domain Adaptation for Audio-visual Speech Enhancement

Chenxi Wang[1], Hang Chen[1], Jun Du[1,*], Chenyue Zhang[1], Yuling Ren[2], Qinglong Li[2], Ruibo Liu[2], and Chin-Hui Lee[3]

[1] University of Science and Technology of China, China
[2] China Mobile Online Services Company Limited, China
[3] Georgia Institute of Technology, USA
jundu@ustc.edu.cn

Abstract. This paper presents the DA-AVSE system developed for the ASRU 2023 Audio-Visual Speech Enhancement (AVSE) Challenge. We initially employed three well-established AVSE models: MEASE, MT-MEASE, and PLMEASE. These models demonstrated effectiveness even without utilizing matched data for training. To further enhance the performance, we introduced a domain adaptation method. More specifically, we utilized pseudo-labels generated by the models above in conjunction with the official baseline to fine-tune each model. Through extensive experiments, we observed that our method significantly improved the models' generalization to the target test set, regardless of whether the training and testing conditions matched. Additionally, we implemented a multi-model fusion strategy to enhance the overall model performance further. Our system exhibited significant improvements in all objective metrics, including PESQ, STOI, and SiSDR, compared to almost all competing teams. As a result, our system ranked the 2nd place in the objective metrics comparison for track 1.

Keywords: Speech enhancement · domain adaptation · audio-visual.

1 Introduction

There are many speech-related applications, such as automatic speech recognition [1], hearing aids [2], video conferencing [3] in reality. But speech is often interfered with by noise in the daily acoustic environment, which has a bad impact on the performance of the applications. Therefore, speech enhancement to reduce noise while maintaining speech quality and intelligibility is important research with great practical application value.

The research on speech enhancement has been developed for decades. Traditional speech enhancement methods are based on statistical signal processing, including spectral subtraction [4], Wiener filtering [5] and minimum mean squared error (MMSE) estimation [6]. In recent years, deep learning has boomed and achieved promising results in speech enhancement [7,8]. Deep learning-based speech enhancement methods recover clean speech from noisy speech via a deep neural network. They can better deal with non-stationary noise in real acoustic scenes than traditional methods. The deep learning-based methods can be divided into the audio-only speech enhancement (AOSE) method and the audio-visual speech enhancement (AVSE) method. It's difficult for the AOSE method to handle scenarios with multiple speakers, but the AVSE method can do better in this case because visual information can assist speech enhancement [9].

Many previous papers [10, 11] have pointed out that the AVSE models bring significant improvements compared to audio-only algorithms, especially in low SNRs. Since the visual cues, such as facial and lip movements, are immune to acoustic conditions. Ephrat et al. [12] proposed an AVSE model for complex ratio mask estimation to separate speech from overlapping speech and background noises. Hou et al. [13] used a multimodal deep convolutional neural network that received both noisy audio and lip images as input and generated enhanced audio and lip images as output, fully leveraging visual information. Chen et al. [14] adopted a multi-modal embedding extracted from a pre-trained articulation place classifier to avoid performance distortion at high SNRs, and realized a significant improvement. [15] designed a lighter model structure by removing the visual feature extraction network with visual data compression, and experimental results confirmed that it provides better performance than the audio-only and AVSE systems without visual data compression. [16] fused audio-visual features layer by layer and introduced the channel and spectral attention mechanisms to pay more attention to informative regions of the fused AV feature maps.

Data-driven speech enhancement methods typically use many noisy and clean speech pairs for training. In the acoustic environment of practical application scenarios, noise types are always not included in the training, which degrades speech enhancement performance. To improve the generalization ability of the speech enhancement models for unseen acoustic environments, the usual practice is to increase the types of acoustic conditions during training as much as possible. However,it is not feasible to include all possible conditions during training, so the mismatch between training and testing of data-driven models has always been existing.

Domain adaptation can be used as a solution to the above problem. It has been extensively studied in computer vision in recent years [17, 18], but it is uncommon in speech signal processing. Domain adaptation aims to transfer the knowledge learned from the source domain to the target domain. Domain adaptation can be used to address the degradation of performance when migrating speech enhancement models to unseen acoustic environments. [19] proposed a domain adversarial training method by utilizing unlabelled target domain noisy speech to extract noise-invariant features, which improved the speech enhancement performance in the unseen target domain. [20] designed a cross-task transfer learning method using paired senone classifiers to align speech signals in the source and target domains and transfer knowledge from the source domain to the target domain through multi-task learning. [21] used a Relativistic Discriminator and Multi-Kernel Maximum Mean Discrepancy (MK-MMD) to align the speech distributions between the source and target domains, thereby improving the speech enhancement model's performance on the target domain.

The Audio-Visual Speech Enhancement (AVSE) Challenge 2023 aims to explore novel approaches to audio-visual speech enhancement. The objective evaluation indicators of the challenge include PESQ, STOI and SISDR. The challenge is divided into two tracks: track 1, which prohibits the use of additional data, and track 2, which allows it. We participated in track 1. In our work, we first use the official training set released by the challenge to train three models, namely MEASE [14], MTMEASE [22] and PLMEASE, which is a model that applies the audio-visual progressive learning framework proposed in [23] to MEASE. These three models achieve competitive results on the official evaluation set. To improve the performance of our system on the evaluation set, we propose a simple and efficient domain adaptation method called Multi-Model Mixture Pseudo-Label Domain Adaptation (MMMP-DA). The method jointly uses pseudo-labels generated by multiple models to fine-tune the models. Specifically, we generate pseudo-labels for the unlabeled evaluation set using three trained models and the officially trained baseline [24]. Then a small-scale dataset is simulated with pseudo-labels to fine-tune the four models. The experimental results demonstrate that our MMMP-DA method can effectively leverage the complementarity among multiple models and enhance their adaptation ability to the evaluation set, regardless of whether the training and evaluation environments match.

Furthermore, we use a multi-model fusion strategy to fuse the enhanced speech predicted by all models, reducing the prediction bias of individual models and resulting in more robust and reliable predictions. After applying our MMMP-DA method and the multi-model fusion strategy, our DA-AVSE submission system achieves 1.77 PESQ, 71.23% STOI and 7.68 SISDR on the evaluation set of the AVSE Challenge 2023, which ranked 2nd in track 1.

2 Proposed system

2.1 Employed Model

We use four models in the challenge: the official baseline model, MEASE, MT-MEASE and PLMEASE. The input of the official baseline model includes the magnitude spectrum and the video frames, and the output of the model is the mask of the magnitude spectrum. The MEASE model was proposed in [14], whose full name is multimodal embedding aware speech enhancement model. Its structure diagram is shown in Figure 1. The MEASE includes a multimodal embedding extractor and an embedding-aware speech enhancement network. There are two branches inside the multimodal embedding extractor, which extract audio embedding and visual embedding respectively. The input of the audio branch is the Mel Filter Bank (FBANK) feature of noisy speech, and the audio embedding is obtained through a 1D convolutional layer, a batch normalization, a ReLU activation and an 18-layer variant of ResNet [25] in which all 2D convolutions are replaced with 1D convolutions. The input of the visual branch is the lip frames cropped from the video, and the visual embedding is obtained through a 3D convolutional layer, a batch normalization, a ReLU activation, a 3D max-pooling layer and an 18-layer ResNet. Then the audio and visual embeddings are concatenated over channel dimension and fused through the fusion module containing a 2-layer Bidirectional Gated Recurrent Unit (BiGRU) to obtain the multimodal embedding.

The 1D ConvBlock in the embedding-aware speech enhancement network includes a 1D convolution layer with a residual connection, a ReLU activation, and a batch normalization. The embedding-aware speech enhancement network takes a concatenation of multimodal embedding and noisy log power spectrum (LPS) [26] as input and predicts the ideal ratio mask (IRM) [27] of the noisy speech.

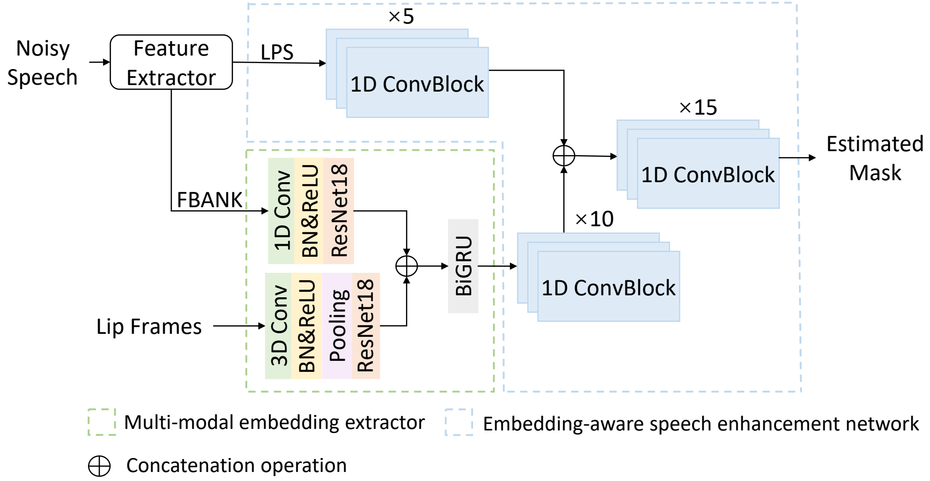


Fig. 1. Schematic diagram of MEASE.

The multimodal embedding extractor is pretrained with the articulation place classification task [14] and kept frozen during the embedding-aware speech enhancement network training. The forward propagation process of the enhancement network for a given sample (A, V), where A represents the noisy speech and V represents the corresponding video, is represented by Eq. (1), where W refers to the model parameters. The enhancement network is trained by minimizing the mean square error (MSE) loss between the predicted ideal ratio mask $\dot{M}$ and the target ideal ratio mask M, as shown in Eq. (2). During training, the network's parameters are updated by backpropagating the loss, which is represented by Eq. (3) and involves the learning rate lr.

$$\dot{M} = \mathcal{F}(A, V, W) \tag{1}$$

$$L_{\text{training}} = \sum \left\| \dot{M} - M \right\|_2^2 \tag{2}$$

$$W \leftarrow W - lr\frac{\partial \mathcal{L}_{\text{training}}}{\partial W} \tag{3}$$

Building on the work of MEASE, [22] considered the differences between audio and visual modalities. Specifically, audio contains more and finer acoustic information compared to video, which is less and rougher. Therefore, to prevent a loss of acoustic details, a finer classification task was used to extract audio embeddings. The phone label was found to be more suitable than the articulation place label due to its finer granularity. Expanding on this idea, [22] proposed a multi-task pre-training method for the multimodal embedding extractor, using both the phone and articulation place labels as training targets to extract more effective multimodal embeddings. The MEASE model that applies this multi-task pre-training method is called multi-task multimodal embedding aware speech enhancement (MTMEASE). The forward propagation and loss backpropagation methods of the enhancement network in the MTMEASE are the same as those in the enhancement network of the MEASE.

In [23], a mask-based audio-visual progressive learning speech enhancement (AVPL) model was proposed to address the problem of a large signal-to-noise ratio (SNR) gap between the learning target and input noisy speech. The model accomplishes this by dividing the mapping between noisy and clean speech into multiple stages, gradually narrowing the SNR gap at each stage. The first stage of AVPL takes a concatenation of the pre-trained visual embedding and the noisy LPS feature as input. In each subsequent stage, the visual embedding and the representation outputted by the previous stage are used as input. The final stage of AVPL predicts the IRM of the noisy speech. The AVPL is able to suppress more noise while preserving more spectral information. We replaced the input visual embeddings of AVPL with multimodal embeddings, and named the resulting model the progressive learning multimodal embedding aware speech enhancement (PLMEASE) model. The motivation for this change is to take advantage of the complementarity of audio and video modalities in multimodal embedding to further improve the performance of speech enhancement. The forward and backward propagation processes of PLMEASE are the same as those of AVPL described in [23].

2.2 Proposed Multi-Model Mixture Pseudo-Label Domain Adaptation (MMMP-DA) method

Our method is model-agnostic and can be applied to multiple models. We regard the labeled training set as the source domain and the unlabeled evaluation set as the target domain. The models are trained on the source domain before applying our MMMP-DA method. The structure of the MMMP-DA method is shown in Figure 2. Assuming a total of K AVSE models, the evaluation set has Q unlabeled noisy speech and the corresponding videos. We use $A = \{A_i, 0 \leq i < Q\}$ to represent the noisy speech feature set of the evaluation set, and $V = \{V_i, 0 \leq i < Q\}$ to represent the visual feature set. The steps of our method are summarized as follows:

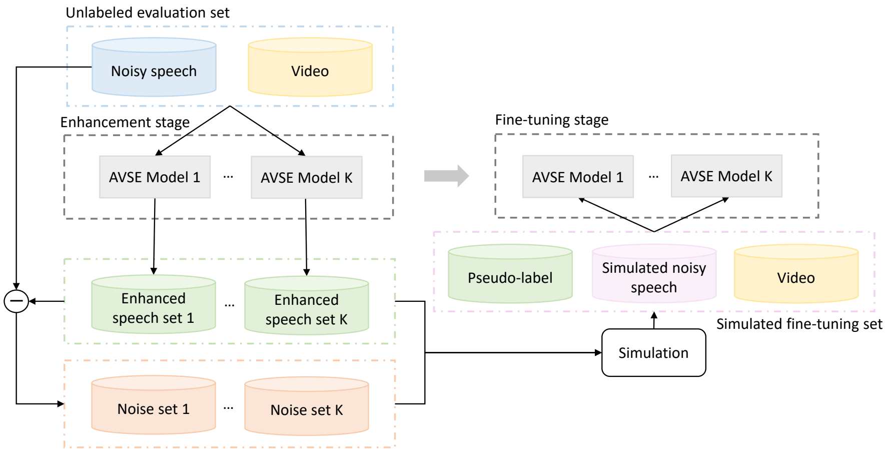


Fig. 2. The proposed MMMP-DA method.

Step 1: The first step is the enhancement stage. The K trained source AVSE models predict pseudo-labels on the target unlabeled evaluation set, respectively. For the i-th sample (A_i, V_i), the predicted mask of the j-th model $\dot{M}_i^j$ can be obtained using Eq. (4), where W_j represents the parameters of the j-th model.

$$\dot{M}_i^j = \mathcal{F}^j(A_i, V_i, W_j) \tag{4}$$

The enhanced speech feature and noise feature of the i-th sample generated by the j-th model can be obtained by Eq. (5) and Eq. (6), respectively.

$$\hat{E}_i^j = A_i \odot \dot{M}_i^j \tag{5}$$

$$\hat{N}_i^j = A_i \odot (1 - \dot{M}_i^j) \tag{6}$$

Each model can get an enhanced speech feature set and a noise feature set of evaluation set. We collect all the enhanced speech feature sets from the K models into a set represented in Eq. (7), and we collect the corresponding noise feature sets into another set represented as shown in Eq. (8).

$$\hat{E} = \left\{\hat{E}_0^0, \cdots, \hat{E}_{Q-1}^0, \cdots, \hat{E}_0^{K-1}, \cdots, \hat{E}_{Q-1}^{K-1}\right\} \tag{7}$$

$$\hat{N} = \left\{\hat{N}_0^0, \cdots, \hat{N}_{Q-1}^0, \cdots, \hat{N}_0^{K-1}, \cdots, \hat{N}_{Q-1}^{K-1}\right\} \tag{8}$$

Step 2: The second step is random simulation, where we create a dataset with a total duration of 20 hours. For each sample, we randomly select an enhanced speech feature $\hat{E}_i^j$ from the set $\hat{E}$ and a noise feature $\hat{N}_u^v$ from the set $\hat{N}$. We then randomly select a value from the set signal-to-noise ratio (SNR) range as the SNR and use it to calculate the noise adjustment factor α during the simulation.We denote the simulated sample as $\tilde{A}_i^u$, and the simulation formula is shown in Eq. (9). The corresponding visual feature is V_i. The label of this sample can be calculated with Eq. (10).

$$\tilde{A}_i^u = \hat{E}_i^j + \alpha \hat{N}_u^v \tag{9}$$

$$\tilde{M}_i^u = \frac{\left\|\hat{E}_i^j\right\|_2^2}{\left\|\hat{E}_i^j\right\|_2^2 + \left\|\alpha \hat{N}_u^v\right\|_2^2} \tag{10}$$

Step 3: The third step involves fine-tuning models using the simulated dataset separately. During the fine-tuning stage, we employ the k-fold cross-validation method [28] which divides the simulated dataset into k subsets for a total of k iterations. We leave one fold for testing each iteration and use the remaining k-1 folds to train the model. This method allowed us to fully use the data, avoid over-fitting issues caused by insufficient data, and improve the model's generalization ability. When the input is the simulated sample $(\tilde{A}_i^u, V_i)$, the output mask $\hat{M}_i^j$ of the j-th model is calculated using Eq. (11).

$$\hat{M}_{i,u}^j = \mathcal{F}^j(\tilde{A}_i^u, V_i, W_j) \tag{11}$$

The loss function we use during fine-tuning is the MSE loss between the predicted mask and the target mask, as shown in Eq. (12). The update formula of the model parameters is shown in Eq. (13), where lr is the learning rate.

$$L_{\text{finetuning}} = \sum \left\|\hat{M}_{i,u}^j - \tilde{M}_i^u\right\|_2^2 \tag{12}$$

$$W \leftarrow W - lr\frac{\partial \mathcal{L}_{\text{finetuning}}}{\partial W} \tag{13}$$

3 Experiments

3.1 Datasets

The AVSE Challenge 2023 provides 113 hours of training set and 8.5 hours of development set. Audio tracks of interferers are composed of a single competing speaker or a noise source in the following ranges: -15 dB to 5 dB (competing speaker) and -10 dB to 10 dB (noise) [24]. The videos of the target speakers and the competing speakers in the training set are selected from the LRS3 dataset [29]. Noise data mainly comes from Clarity Challenge (First edition) [30], Freesound [31], and DNS Challenge (Second edition) [32]. All audio files are monaural speech with a 16 kHz sampling frequency and 16 bits of bit depth. The target speakers, competing speakers and noise files of the training and development set are all disjoint but share the same noise categories.

The official evaluation set has 1,389 extracted sentences from 30 speakers. Approximately half of the mixed speech in the evaluation set has a competing speaker scenario while the other half has noise. There are six competing speakers.

The noise types used in the evaluation set are a subset of the noise types used in the training and development sets.

To evaluate the generalization ability of our MMMP-DA method, we create an acoustic environment that does not match the evaluation set. Specifically, we build an audio-visual dataset based on the benchmark released by the Multimodal Information-based Speech Processing (MISP) 2021 Challenge [33]. The MISP2021 Challenge Audio-Visual Speech Recognition (AVSR) dataset contains 122.53 hours of audio-visual data recorded in a real-home TV room, with the language being Chinese. We utilize near-field audio and corresponding mid-field video from this dataset and mix it with MISP home-scene noise at 6 levels of SNRs (-15 dB, -10 dB, -5 dB, 0 dB, 5 dB, and 10 dB) to create a 132-hour training set and a 10-hour development set. We named this dataset as the MISP home-noise dataset. The acoustic environment of this dataset is significantly mismatched from that of the AVSE Challenge 2023 evaluation set, as the noise and language types are completely different.

3.2 Experimental settings

For the MEASE, MTMEASE, and PLMEASE models, we utilize a 25-ms Hanning window and an overlap of 10 ms to extract audio frames during the audio preprocessing phase. The spectra of each frame is computed by a 400-point short-time Fourier transform. The videos are resampled to 25 fps and the lip area with a size of 98×98 pixels is cropped from each video frame. We compute 201-dimensional LPS features as the input of the enhancement network, 40-dimensional FBANK features and lip frames as input to the multimodal embedding extractor.

The objective metrics used on the challenge leaderboard include perceptual evaluation of speech quality (PESQ) [34], short-time objective intelligibility (STOI) [35] and Scale-Invariant Signal-to-Distortion Ratio (SISDR) [36]. Higher is better on three metrics.

The training methods of the models MEASE, MTMEASE and PLMEASE are similar. Firstly, we train the multimodal embedding extractor for 100 epochs with the Adam optimizer. The initial learning rate is 3e-4, decreasing on log scale after 30 epochs. After training the multimodal embedding extractor, we freeze it and train the speech enhancement network for 100 epochs using the Adam optimizer. The initial learning rate is set to 1e-3 and halved if the validation loss does not decrease for three epochs.

During random simulation, we generate a dataset of 20 hours of noisy-clean pairs with SNRs ranging from -10dB to 0dB. In the fine-tuning stage, the number of cross-validation folds is set to 10. We use the Adam optimizer with an initial learning rate of 3e-4, which is halved if the validation loss fails to improve after three consecutive epochs.

3.3 Results of the MEASE, MTMEASE and PLMEASE in matched scenario

In this section, we present the results of the MEASE, MTMEASE and PLMEASE models on the evaluation set after being trained on the official training set, which represents a matched condition between training and test data. As evidenced quantitatively in Table 1, the three aforementioned models demonstrated competitive performance across all evaluation metrics.

Table 1. Performance of MEASE, MTMEASE and PLMEASE on the evaluation set in the matched scenario.

Model	PESQ	STOI(%)	SISDR
Noisy	1.14	44.10	-5.07
Baseline	1.41	55.63	3.67
MEASE	1.60	67.66	5.34
MTMEASE	1.61	67.86	5.46
PLMEASE	1.56	66.28	4.97

3.4 Performance analysis of the proposed MMMP-DA method

To demonstrate the effectiveness of our proposed method, we compare the performance of our MMMP-DA method with a common single-model pseudo-label domain adaptation method, denoted as SMP-DA. The SMP-DA method only uses pseudo-labels and noise generated by a single model to simulate the fine-tuned dataset, which means that domain adaptation is only performed on a single model. We apply the SMP-DA and proposed MMMP-DA methods on four models: the official baseline, MEASE, MTMEASE and PLMEASE. The results are shown in Table 2.

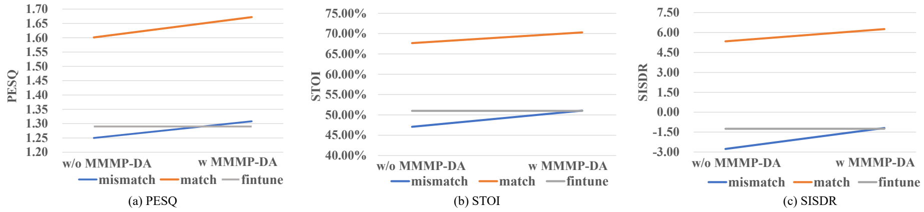
Table 2. Performance comparison of SMP-DA and MMMP-DA methods on the baseline, MEASE, MTMEASE and PLMEASE models on the evaluation set in matched scenario.

Model	SMP-DA			MMMP-DA		
	PESQ	STOI(%)	SISDR	PESQ	STOI(%)	SISDR
Baseline	1.60	64.26	8.01	1.63	65.51	8.43
MEASE	1.64	69.09	5.79	1.68	70.30	6.26
MTMEASE	1.65	69.25	5.85	1.68	70.30	6.24
PLMEASE	1.61	68.54	5.67	1.64	69.25	5.93

By comparing the results presented in Tables 1 and 2, we can see that our MMMP-DA method consistently improves the performance of all models, indicating that our method is robust and effective. For example, the SISDR of the baseline model increased from 3.67 to 8.43, and the STOI of the PLMEASE model increased from 66.28% to 69.25%.

Comparing the results of the SMP-DA and MMMP-DA methods in Table 2, it is clear that the MMMP-DA method consistently outperforms the SMP-DA method. These findings suggest that leveraging prior knowledge from multiple models through the MMMP-DA method is more effective in facilitating domain adaptation than using a single model, as in the SMP-DA method. Specifically, the MMMP-DA method can leverage the complementarity among multiple models and capture a wider range of variations in the target evaluation set, enabling effective domain adaptation across multiple models. In contrast, using a single model for domain adaptation may lead to limited effectiveness in improving its performance due to the model's inability to capture all the variations in the target evaluation set.

To demonstrate the validity of our MMMP-DA method in mismatched scenarios, we evaluate the performance of the MEASE model with and without the MMMP-DA method applied in a mismatched scenario using the MISP home-noise dataset as the training set. We also set up a fine-tuning scenario training with the MISP home-noise dataset and fine-tuning on the official development set to compare with the MMMP-DA method. The number of cross-validation folds when fine-tuning with the development set and the learning rate setting are the same as that of our MMMP-DA method. We present the performance comparison on the evaluation set in Figure 3.



(a) PESQ (b) STOI (c) SISDR

Fig. 3. The performance comparison of MEASE on the evaluation set. "w/o MMMP-DA" stands for MEASE without applying our MMMP-DA method. "w MMMP-DA" stands for MEASE applying our MMMP-DA method. "mismatch" represents that MEASE is trained in the mismatched scenario, that is, the MISP home-noise dataset. "match" represents that MEASE is trained in the matched scenario which is the official training set. "finetune" means that the MEASE is fine-tuned with a small amount of matching data after training in mismatched scenario.

Figure 3 shows that the MMMP-DA method steadily improves the performance of the MEASE model in all three evaluation metrics, regardless of whether the environments are matched or mismatched. In the case of mismatched environments, our method enables the model to achieve comparable or even better performance than the model fine-tuned with the matched development set. This suggests that the MMMP-DA method can effectively improve the model's generalization and make it more suitable for the evaluation set, even when labeled real matching data is not available during the fine-tuning stage.

3.5 Objective results on AVSE Challenge 2023

In this section, we introduce the method we use in the challenge. We first use the official training set to train the MEASE, MTMEASE and PLMEASE models, and then we jointly apply our MMMP-DA method on these three models and the official trained baseline provided by the challenge. Finally, we fuse the prediction results of the evaluation set from the four models, that is, average the predicted time domain waveforms to obtain the final predicted waveform. This operation can effectively reduce the bias and variance of a single model, thereby improving the stability and accuracy of predictions.

Table 3. Performance comparison of our system and some other competing systems on the evaluation set.

System	PESQ	STOI(%)	SISDR
Baseline	1.41	55.63	3.67
BioASP_CITI	1.41	53.78	3.61
AVSE02	1.61	68.21	8.81
Ict_avsu	1.66	67.67	6.73
Merl	2.71	83.76	14.43
DA-AVSE	1.77	71.23	7.68

Table 3 presents a comparison of the performance of our DA-AVSE system with some other participating systems in the challenge track 1. The table shows that our DA-AVSE system outperforms almost all other competing systems regarding all three evaluation metrics. Specifically, our system achieves a PESQ of 1.77, an STOI of 71.23%, a SISDR of 7.68 dB. These results demonstrate the effectiveness of the MMMP-DA method and the model fusion strategy.

4 Conclusions

This paper presents our submission to the AVSE Challenge 2023 track 1, where we introduced three AVSE models: MEASE, MTMEASE, and PLMEASE, yielding competitive results. Additionally, we propose a domain adaptation method called MMMP-DA to enhance the model's generalization capability to the target test set. By applying the MMMP-DA method to the official baseline, as well as the MEASE, MTMEASE, and PLMEASE models, and employing a multi-model fusion strategy, we further improve the overall model performance. Our final results on the evaluation set demonstrate notable achievements, ranked 2nd on track 1 in terms of objective metrics.

References

1. Chia-Yu Li and Ngoc Thang Vu. Improving speech recognition on noisy speech via speech enhancement with multi-discriminators cyclegan. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 830–836, 2021.
2. Issa Panahi, Nasser Kehtarnavaz, and Linda Thibodeau. Smartphone-based noise adaptive speech enhancement for hearing aid applications. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 85–88, 2016.
3. Ross Cutler, Ramin Mehran, Sam Johnson, Cha Zhang, Adam Kirk, Oliver Whyte, and Adarsh Kowdle. Multimodal active speaker detection and virtual cinematography for video conferencing. In *ICASSP 2020*, pages 4527–4531, 2020.
4. S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120, 1979.
5. Jae Lim and A. Oppenheim. All-pole modeling of degraded speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(3):197–210, 1978.
6. Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):443–445, 1985.
7. Yuxuan Wang, Arun Narayanan, and DeLiang Wang. On training targets for supervised speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1849–1858, 2014.
8. Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):7–19, 2015.
9. Shrishti Saha Shetu, Soumitro Chakrabarty, and Emanuël A. P. Habets. An empirical study of visual features for dnn based audio-visual speech enhancement in multi-talker environments. In *ICASSP 2021*, pages 8418–8422, 2021.
10. William H Sumby and Irwin Pollack. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215, 1954.
11. Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.
12. Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.
13. Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):117–128, 2018.
14. 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*, 143:171–182, 2021.
15. Shang-Yi Chuang, Yu Tsao, Chen-Chou Lo, and Hsin-Min Wang. Lite audio-visual speech enhancement. *arXiv preprint arXiv:2005.11769*, 2020.
16. Xinmeng Xu, Y Wang, D Xu, Y Peng, C Zhang, J Jia, Y Wang, and B Chen. Mffcn: multi-layer feature fusion convolution network for audio-visual speech enhancement. *arXiv preprint*, 2021.
17. Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *CoRR*, abs/0902.3430, 2009.

18. Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971, 2017.
19. Chien-Feng Liao, Yu Tsao, Hung-Yi Lee, and Hsin-Min Wang. Noise Adaptive Speech Enhancement Using Domain Adversarial Training. In *Proc. Interspeech 2019*, pages 3148–3152, 2019.
20. Sicheng Wang, Wei Li, Sabato Marco Siniscalchi, and Chin-Hui Lee. A cross-task transfer learning approach to adapting deep speech enhancement models to unseen background noise using paired senone classifiers. In *ICASSP 2020*, pages 6219–6223, 2020.
21. Jiaming Cheng, Ruiyu Liang, Zhenlin Liang, Li Zhao, Chengwei Huang, and Björn Schuller. A deep adaptation network for speech enhancement: Combining a relativistic discriminator with multi-kernel maximum mean discrepancy. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:41–53, 2021.
22. 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)*, pages 255–259, 2022.
23. 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*, pages 1–5, 2023.
24. Andrea Lorena Aldana Blanco, Cassia Valentini-Botinhao, Ondrej Klejch, Mandar Gogate, Kia Dashtipour, Amir Hussain, and Peter Bell. Avse challenge: Audio-visual speech enhancement challenge. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 465–471, 2023.
25. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
26. Jun Du and Qiang Huo. A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions. In *Interspeech*, 2008.
27. Christopher Hummersone, Toby Stokes, and Tim S. Brookes. On the ideal ratio mask as the goal of computational auditory scene analysis. 2014.
28. Sanjay Yadav and Sanyam Shukla. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, pages 78–83, 2016.
29. Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition. *CoRR*, abs/1809.00496, 2018.
30. Simone Graetzer, Jon Barker, Trevor J. Cox, Michael A. Akeroyd, John F. Culling, Graham Naylor, Eszter Porter, and Rhoddy Viveros Muñoz. Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing. In *Interspeech*, 2021.
31. Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings. *Proceedings of Meetings on Acoustics*, 19(1):035081, 06 2013.
32. Chandan K. A. Reddy, Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan. Icassp 2021 deep noise suppression challenge. In *ICASSP 2021*, pages 6623–6627, 2021.

33. Hang Chen, Hengshun Zhou, Jun Du, Chin-Hui Lee, Jingdong Chen, Shinji Watanabe, Sabato Marco Siniscalchi, Odette Scharenborg, Di-Yuan Liu, Bao-Cai Yin, et al. The first multimodal information based speech processing (misp) challenge: Data, tasks, baselines and results. In *ICASSP 2022*, pages 9266–9270. IEEE, 2022.
34. 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)*, volume 2, pages 749–752 vol.2, 2001.
35. Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.
36. E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.