AUDIO-VISUAL SPEECH ENHANCEMENT AND SEPARATION BY UTILIZING MULTI-MODAL SELF-SUPERVISED EMBEDDINGS

I-Chun Chern[1]*, Kuo-Hsuan Hung*[2,3]*, Yi-Ting Chen*[3]*, Tassadaq Hussain*[4]
Mandar Gogate[4]*, Amir Hussain*[4]*, Yu Tsao*[3]*, Jen-Cheng Hou*[3]

[1]Carnegie Mellon University, USA [2]National Taiwan University, Taiwan
[3]Academia Sinica, Taiwan [4]Edinburgh Napier University, UK

ABSTRACT

AV-HuBERT, a multi-modal self-supervised learning model, has been shown to be effective for categorical problems such as automatic speech recognition and lip-reading. This suggests that useful audio-visual speech representations can be obtained via utilizing multi-modal self-supervised embeddings. Nevertheless, it is unclear if such representations can be generalized to solve real-world multi-modal AV regression tasks, such as audio-visual speech enhancement (AVSE) and audio-visual speech separation (AVSS). In this study, we leveraged the pre-trained AV-HuBERT model followed by an SE module for AVSE and AVSS. Comparative experimental results demonstrate that our proposed model performs better than the state-of-the-art AVSE and traditional audio-only SE models. In summary, our results confirm the effectiveness of our proposed model for the AVSS task with proper fine-tuning strategies, demonstrating that multi-modal self-supervised embeddings obtained from AV-HuBERT can be generalized to audio-visual regression tasks.

Index Terms— Audio-Visual Speech Enhancement, Audio-Visual Speech Separation, AV-HuBERT

1. INTRODUCTION

Speech enhancement (SE) and speech separation (SS) aim to extract speech signals of interest from a given utterance mixed with unwanted audio signals. With recent developments in deep learning (DL), DL-based methods have demonstrated better results than traditional SE and SS methods, either for audio-only or audio-visual (AV) applications [1, 2, 3, 4, 5, 6]. Nevertheless, most DL-based AVSE and AVSS models have their own specific modules designed to better integrate the audio-visual information for the target task, which may not be favorable from the viewpoints of some current DL model design philosophies. One popular learning paradigm is designing a unified scheme that can learn generalizable representations with minor model modifications for different tasks. Self-supervised learning (SSL) is a popular learning strategy for this purpose. SSL uses the data itself for its own learning supervision. It has been effective in several multi-modal applications, such as vision-and-language [7] and audio-visual learning [8, 9]. Specifically, for audio-visual applications [8, 9], a transformer model [10] is pre-trained with audio-visual data and then fine-tuned for classification tasks, such as automatic speech recognition (ASR) and lip-reading.

In this study, we investigated whether pre-trained audio-visual models are beneficial in regression-based speech processing tasks, namely AVSE and AVSS. In particular, we consider AV-HuBERT [8] our audio-visual SSL model and combine AV-HuBERT with a simple neural regression model for AVSE and AVSS tasks. The results show that model performances can be improved, indicating that pre-trained audio-visual representations from AV-HuBERT are beneficial in audio-visual regression tasks.

The paper is organized as follows. Section 2 gives some review on related works. Section 3 presents the details of the proposed approach, and Section 4 demonstrates our experiments and results. A summary of the paper is presented in Section 5.

2. RELATED RESEARCH

Recently, an increasing number of audio-only self-supervised models have been applied to SE and SS. For applications in SE, some studies [11, 12, 13] utilized pretrained latent representations as the SE model input and boosted the performance by incorporating cross-domain features [14]. For applications in SS, the authors of [15] proposed a self-supervised pre-training approach to stabilize label assignments when training SS models.

There are relatively few studies on self-supervised pre-training for audio-visual regression-based tasks [8, 9]. AV-HuBERT [8] is an extension of the Hubert [16] model (an audio-only SSL model) for multimodal pre-training. The authors in [17] leveraged pre-learned representations from a pretrained AV-HuBERT for speaker classification and verification tasks. In contrast, we focused on two regression-based speech processing tasks, namely AVSE and AVSS.

* Corresponding author: Jen-Cheng Hou, jchou@citi.sinica.edu.tw

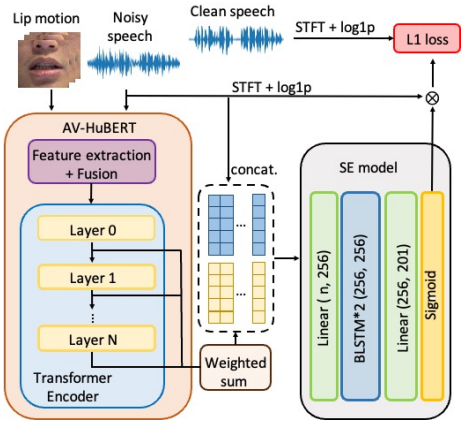

Fig. 1. Our proposed AVSE model is based on the AV-HuBERT, followed by a SE model.

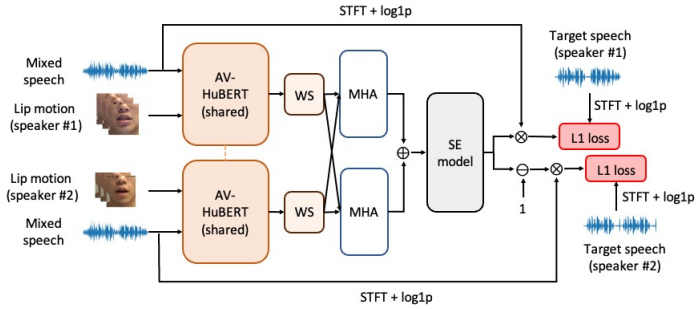

Fig. 2. Proposed audio-visual SS framework. The output features of the shared AV-HuBERT from the two speakers are coupled via cross-attention and then are fed into the neural regression model for SS. The WS and MHA blocks represent the weighted-sum and the multi-head attention modules, respectively.

3. PROPOSED METHOD

3.1. Audio-Visual SE Model

The proposed AVSE approach leverages the AV-HuBERT model as an upstream processor, as shown in Fig. 1. The lip image sequence $V_{1:T}$ and noisy speech $A^n_{1:T}$ are fed into the AV-HuBERT; the representations from each layer of the transformer encoder are denoted as H^l, where $0 \leq l \leq N$, and N is the number of layers. Inspired by [11, 12], a trainable function $w(\cdot)$ is applied to the representations from all layers as follows:

$$H_{\text{WS}} = \sum_{l=0}^{L} w(l)\, H^{(l)} \tag{1}$$

where $w(l)$ denotes the weight of the l-th layer and has the properties $w(l) \geq 0$ and $\sum_l w(l) = 1$. H_{WS} are then concatenated with the $log1p$ spectrogram feature from the noisy speech. The concatenated features are subsequently fed into a neural SE module consisting of fully connected layers (FC) and a two-layer bidirectional long short-term memory module (BLSTM). The output of the SE module is a soft mask and is multiplied by the magnitude of the spectrogram of the noisy speech. The training objective is to minimize the $L1$ distance between the multiplied spectrogram and that generated from the clean speech.

3.2. Audio-Visual SS Model

Similar to the proposed AVSE model, our proposed AVSS model uses the AV-HuBERT model as the front-end module to process audio-visual inputs. Fig. 2 shows the overall architecture of the AVSS system. The two image sequences for the speaker lip movements in a two-speaker video are denoted as $V^{s1}_{1:T}$ and $V^{s2}_{1:T}$. The mixed speech is expressed as $A^m_{1:T}$. A shared AV-Hubert model is used to generate the multi-modal representations for each speaker, followed by a weighted-sum operation, which is the same as Eq.1. The resulting outputs are denoted as $H_{\text{WS,sp1}}$ and $H_{\text{WS,sp2}}$. To better couple the multi-modal features obtained from the two speakers, we applied cross-attention over $H_{\text{WS,sp1}}$ and $H_{\text{WS,sp2}}$ with a layer of multi-head attention (MHA) mechanism, which is expressed as:

$$\begin{aligned} O_{sp1} &= MHA(H_{\text{WS,sp1}}, H_{\text{WS,sp2}}) \\ O_{sp2} &= MHA(H_{\text{WS,sp2}}, H_{\text{WS,sp1}}) \end{aligned} \tag{2}$$

The outputs of the MHA modules are then summed and fed into the SE module, which has the same architecture as the previous AVSE model. The objective of the loss is to minimize the $L1$ distances between the masked spectrograms and target spectrograms, which can be expressed as:

$$\begin{aligned} L_{avss} &= dist(SE(O_{sp1} \bigoplus O_{sp2}) \bigotimes Sm, Ssp1) \\ &+ dist((1 - SE(O_{sp1} \bigoplus O_{sp2})) \bigotimes Sm, Ssp2) \end{aligned} \tag{3}$$

where SE, Sm, and S_{sp1} and S_{sp2} denote the SE model, magnitude of the spectrograms of the mixed speech, and speech from each speaker, respectively.

4. EXPERIMENTS

4.1. Dataset and Training strategies

This section presents the experimental setup and results. We evaluate our proposed models for the AVSE and AVSS tasks based on the TSMV dataset [1]. The dataset contains video recordings of 18 native speakers (13 males and 5 females),

[1] https://bio-asplab.citi.sinica.edu.tw/Opensource.html

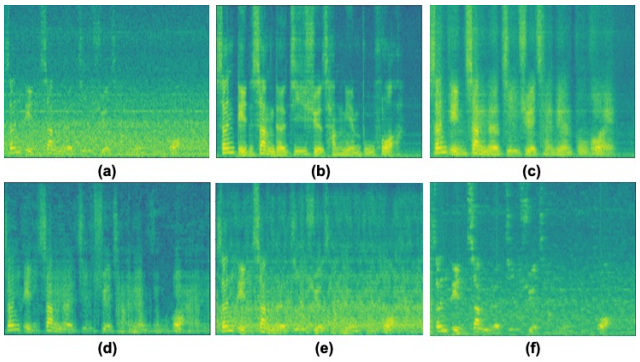

Fig. 3. Spectrograms of the enhanced speech by different methods. (a) Noisy speech of engine noises at -1 dB. (b) Clean speech. (c)-(f) represent enhanced speech by our proposed method, LAVSE, AV-CVAE, and LogMMSE, respectively.

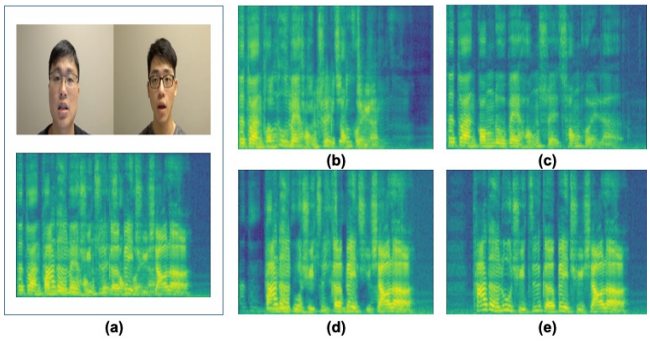

Fig. 4. Results of our proposed AVSS framework. (a) A two-speaker talking video and the spectrogram of the mixed speech. (b) and (c) are spectrograms of the recovered speech for the left speaker and ground truth. (d) and (e) are spectrograms of the recovered speech for the right speaker and ground truth.

each speaking 320 utterances of Mandarin sentences. Each sentence consists of 10 Chinese characters, and the approximate duration for each utterance is 2–4 seconds.

We optimized the AVSE and AVSS models with several training strategies. The first is partial fine-tuning (PF), which means that the weights of the feature extraction parts of AV-HuBERT are fixed, whereas the weights in the SSL block, that is, the transformer encoder, are updated based on the pretrained checkpoint. Second, we investigate the entire fine-tuning (EF), whose difference from PF is the inclusion of weight updating for the feature extraction parts. Training from scratch (TFS) and training of the model without fine-tuning (WF) AV-HuBERT were also performed for comparison. The following sections detail the setup for the experiments for AVSE and AVSS, respectively.

4.2. Audio-Visual SE Task

4.2.1. Experimental setup

For the pre-processing parts in the videos, we cropped the mouth of the speakers via pre-trained CNN detectors [18]. For the audio components, we followed the setup described in [4]. Of the 320 utterances for each speaker, we use the first 200 ones for training, and the remaining 120 ones are used for testing. To form clean-noisy speech pairs for training, the utterances were artificially corrupted by 100 types of noise [19] at five different signal-to-noise ratios (SNRs) ranging from -12 to 12 dB with an increment of 6 dB. The process generates approximately 600 hours of noisy utterances, and 12,000 noisy utterances are randomly selected to form a 9-hour training set to reduce the computation cost. To form the testing set, we selected five types of noise, including baby crying sounds, engine noise, pink noise, music noise, and street noise, with SNRs at -1 dB, -4 dB, -7 dB, and -10 dB.

4.2.2. Model specification and training setup

We used Base AV-HuBERT with 12 transformer layers pretrained on the LRS3 dataset [20] for five iterations as the checkpoint to build our AVSE system. The feature dimensions of the BLSTM model following AV-HuBERT were 256. For the fine-tuning process, the initial learning rate was set at 1e-4, and AdamW [21] was used as the learning optimizer. For each training strategy, we trained the model for 50 epochs and selected the model that yielded the best performance on the validation sets.

4.2.3. Experimental results

In this section, we compare the experimental results of the proposed AVSE model with those of other baseline SE models, including two AVSE models [4, 22] and one traditional audio-only method, namely LogMMSE [23]. We conducted an objective comparison with two standardized evaluation metrics that are widely used to evaluate SE performance—the perceptual evaluation of speech quality (PESQ) [24] and short-time objective intelligibility measure (STOI) [25]. PESQ measures the quality of processed speech, whereas STOI is developed for evaluating speech intelligibility. PESQ and STOI scores range from -0.5 to 4.5 and 0 to 1, respectively. Higher scores indicate better performance.

Table 1 lists the PESQ and STOI scores for each SE method. From the table, we observe that the SE task is quite challenging because the traditional audio-only methods did not yield much improvement. In addition, the models using audio-visual information outperformed the audio-only methods, confirming the effectiveness of including visual formation in SE. Next, focusing on our proposed method, we can see that PF, EF, and WF all outperform TFS, indicating that the learned features in the pretrained AV-HuBERT model

indeed result in a better AVSE system. We also implemented an AVSE system that used only the encoder part of the AV-HuBERT model (i.e., without the transformer part). These results are also reported in Table 1 and termed AVSE (w/o Trans.). From the table, we noticed that AVSE (w/o Trans.) does not perform as well as TFS does, suggesting that the transformer module is also effective for feature learning in the AVSE.

Fig.3 shows spectrograms of the enhanced speech from different SE approaches. We can observe the high-frequency components of the speech enhanced by partially fine-tuning the AV-HuBERT are better preserved than those from the LAVSE and AV-CVAE approaches, showing the consistency of the superiority of our approach.

Methods	Mod.	PESQ	STOI
Noisy		1.18	0.60
LogMMSE [23]	(A)	1.21	0.61
AV-CVAE [22]	(AV)	1.34	0.63
LAVSE [4]	(AV)	1.31	0.61
AVSE (w/o Trans.)	(AV)	1.25	0.61
AVSE (WF)	(AV)	1.30	0.63
AVSE (TFS)	(AV)	1.26	0.60
AVSE (EF)	(AV)	1.37	0.66
AVSE (PF)	(AV)	**1.40**	**0.68**

Table 1. Results of speech enhancement via the audio-only or audio-visual methods.

4.3. Audio-Visual Speech Separation

4.3.1. Experimental setup

AVSS experiments were based on the TMSV dataset. As shown in Fig.4(a), we made a two-talker video by merging two individual videos side by side, while redoing the soundtrack by mixing the original ones. The spectrogram depicted in Fig.4(a) is the mixed speech from the utterances of the left speaker, as shown in Fig.4(c), and the right speaker, as shown in Fig.4(e). We randomly selected two utterances from two different speakers to form a mixed speech for separation. A total of 12,000, 1,200, 1,200 mixed utterances for training, validation and testing were made with spoken sentences and speakers mismatched.

4.3.2. Model specification and training setup

We used the same AV-HuBERT checkpoint as that used for the AVSE task. For the multi-head attention modules, we used 12 heads with a feature dimension of 512. The neural regression module of the AVSS model was the same as that used in the AVSE model. For the fine-tuning process, the initial learning rate, optimizer, and training epochs were the same as those for the AVSE task.

4.3.3. Experimental results

Fig.4 shows spectrograms of the mixed speech and the separated speech by the proposed AVSS approach. Fig.4(a) demonstrates a two-speaker sample video where the speakers are speaking simultaneously. The spectrogram of the mixed speech in Fig.4(a) is composed of the clean utterances from the left speaker and the right speaker. The spectrograms of the separated speech obtained by fine-tuning AV-HuBERT of our AVSS model are presented in Figs.4 (b) and (d), with the respective ground-truth Figs.4 (c) and (e). We can observe from the spectrograms that our method can effectively separate mixed speech using the corresponding visual information.

To quantitatively assess the results of different training strategies, we used two metrics for evaluation: the scale-invariant signal-to-noise ratio (SI-SNR) and source-to-distortion ratio (SDR). These two metrics are common for evaluating separated speech [26]. Table 2 reports the two scores of the separated speech using the different training strategies of our AVSS methods. The scores related to PF were the best among the training strategies, suggesting that the learned representations of AV-HuBERT can be effective for the AVSS task as well. Nevertheless, note that in the AVSS task, TFS is only second to PF, unlike AVSE, where the performance of TFS is worse than that of PF, EF, and WF, as reported in Table.1. We argue that including multiple attention heads improves the coupling of multimodal features from different speakers, thereby closing the performance gap between TFS and the others via leveraging the pre-trained AV-HuBERT features.

Methods	SISNR (dB)	SDR (dB)
AVSS (WF)	3.03	4.08
AVSS (TFS)	3.53	**4.63**
AVSS (EF)	3.27	4.31
AVSS (PF)	**3.59**	4.59

Table 2. Results of audio-visual speech separation by different fine-tuning strategies.

5. CONCLUSION

In this study, we proposed novel AVSE and AVSS frameworks that leveraged a pretrained AV-HuBERT model. For AVSE, we demonstrated that after partially fine-tuning the AV-HuBERT model, our AVSE system outperformed other baseline SE models. For AVSS, we noted similar trends and the advantages of using AV-HuBERT embeddings. In summary, this study demonstrated how a pre-trained AV-HuBERT model can improve the training of AVSE and AVSS tasks, showing its promising ability for AV regression tasks.

6. REFERENCES

[1] Y. Xu et al., "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[2] X. Lu et al., "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech 2013*.

[3] J.-C. Hou et al., "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, Apr. 2018.

[4] S.-Y. Chuang, H.-M. Wang, and Y. Tsao, "Improved lite audio-visual speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1345–1359, 2022.

[5] R. Gao and K. Grauman, "VisualVoice: Audio-visual speech separation with cross-modal consistency," in *Proc. CVPR*, 2021.

[6] J. Lee et al., "Looking into your speech: Learning cross-modal affinity for audio-visual speech separation," in *Proc. CVPR*, 2021.

[7] J. Lu et al., "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. NeurIPS*, 2019.

[8] B. Shi et al., "Learning audio-visual speech representation by masked multimodal cluster prediction," *arXiv preprint arXiv:2201.02184*, 2022.

[9] D. M. Chan et al., "Multi-modal pre-training for automated speech recognition," in *Proc. ICASSP*, 2022.

[10] A. Vaswani et al., "Attention is all you need," in *Proc. NeurIPS*, 2017.

[11] Z. Huang et al., "Investigating self-supervised learning for speech enhancement and separation," in *Proc. ICASSP*, 2022.

[12] H.-S. Tsai et al., "SUPERB-SG: Enhanced speech processing universal PERformance benchmark for semantic and generative capabilities," in *Proc. ACL*, 2022.

[13] Aswin Sivaraman and Minje Kim, "Efficient personalized speech enhancement through self-supervised learning," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–15, 2022.

[14] K.-H. Hung et al., "Boosting self-supervised embeddings for speech enhancement," in *Proc. Interspeech*, 2022.

[15] S.-F. Huang et al., "Stabilizing label assignment for speech separation by self-supervised pre-training," in *Proc. Interspeech*, 2021.

[16] W.-N. Hsu et al., "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[17] B. Shi, A. Mohamed, and W.-N. Hsu, "Learning lip-based audio-visual speaker embeddings with AV-HuBERT," in *Proc. Interspeech*, 2022.

[18] "Dlib face detector library," `http://dlib.net/files/mmod_human_face_detector.dat.bz2`.

[19] G. Hu, "100 nonspeech environmental sounds," `http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html`.

[20] 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.

[21] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[22] Mostafa Sadeghi, Simon Leglaive, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud, "Audio-visual speech enhancement using conditional variational auto-encoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1788–1800, 2020.

[23] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 857–869, Sept. 2005.

[24] A. W. Rix et al., "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001.

[25] C. H. Taal et al., "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, sep 2011.

[26] Yi Luo and Nima Mesgarani, "TaSNet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. ICASSP*, 2018.