AVHuMAR: Audio-Visual Target Speech Extraction with Pre-trained AV-HuBERT and Mask-And-Recover Strategy

Wenxuan Wu, Xueyuan Chen, Xixin Wu, Haizhou Li, Helen Meng
The Chinese University of Hong Kong

1. Introduction

The quest for the understanding of human's cocktail party effect and the implementation of automatic speech separation has never stopped. Recently, speech separation systems like [6, 7] achieved remarkable performance, benefiting downstream tasks such as speech recognition [19] and speech translation [8]. However, these systems have limitations, requiring the number of speakers in advance and sometimes facing the speaker permutation problem [6]. Target speaker extraction (TSE) seeks to extract target speaker's voice from a mixture conditioned on certain cues [20]. Such a task is more consistent with the human selective auditory attention process in a cocktail party [11].

For effective TSE, the choice of proper speaker cues remains a topic of study. Some studies explore the use of target speech clips as speaker cues [21, 18, 5]. However, these techniques suffer from mismatches like different recording scenarios and intra-speaker speech variety [20]. To address these bottlenecks, more stable and noise-invariant visual cues are utilized. AV-ConvTasNet employs a pre-trained visual speech recognition (VSR) encoder to extract target speaker features [17]. MuSE introduces an additional audio encoder to verify target speaker [10]. Nevertheless, practical scenarios may have lip occlusion issues [4], such as camera movements or speakers turning around, altering their frontal view. ImagineNET [12] tackles this by using extra visual refiner blocks to compensate for missing visual cues. Considering the noise and reverberation invariant properties as well as the advanced solutions for visual occlusion, lip movements remain the most effective target speaker cue for the TSE systems.

Recently, AV-HuBERT achieved great success in lip-reading tasks, showing its strong ability to capture audio-visual synchronization [15, 1]. To benefit from such robust audio-visual synchronization knowledge, we integrate pre-trained AV-HuBERT layers into our TSE system. Furthermore, to facilitate the alignment between audio feature space and visual feature space, a novel Mask-And-Recovery (MAR) strategy has been applied to our TSE system. With the integrated AV-HuBERT layers and additional MAR strategy, we propose the AVHuMAR-TSE system. The contributions of this paper could be summarized in three folds:

- First, we integrate the pre-trained AV-HuBERT layers into our proposed audio-visual TSE system, which is called the AVHuBERT-TSE system. To the best of our knowledge, this is the first attempt to combine the AV-TSE system with the audio-visual foundation model.
- To enhance both intra and inter-modality alignments, we further propose the AVHuMAR-TSE system, which jointly optimizes the pre-trained AVHuBERT-TSE system and the integrated MAR block. Experimental improvements demonstrate the effectiveness of the proposed Mask-And-Recovery (MAR) strategy.
- To verify the effectiveness of the proposed MAR strategy, we experiment with different mask durations for the mixture speech and find the best configuration for our AVHuMAR-TSE system.

2. Method

2.1. AVHuBERT-TSE system

The overview structure of the AVHuBERT-TSE system is shown in Fig 1 (a). The speech encoder, speech decoder, and visual encoder follow the structure in MuSE [10]. The visual adapter follows the structure in *reentry* [11]. The contribution is the speaker extractor, shown in Fig 1 (c). Specifically, the speaker extractor block will be repeated r times, at each speaker extractor block, the intermediate estimated target speech $S_{(t)}^{r-1}$ will align with target visual cue $V_{(t)}^{r-1}$ temporally. The refined visual cue will be obtained after passing through the pre-trained AV-HuBERT layers. Two symmetric Conv1D adapters will be inserted to help the model adapt to the pre-trained weights.

2.2. AVHuMAR-TSE system

Except for the visual cues, the intermediate estimated target speech frames contain rich contextual information and target speaker features, which could also guide the extraction process and reduce the significant reliance on visual cues. Instead of extracting intermediate target speech features explicitly, we propose a novel Mask-And-Recover strategy to learn both inter and intro-modality correlations implicitly.

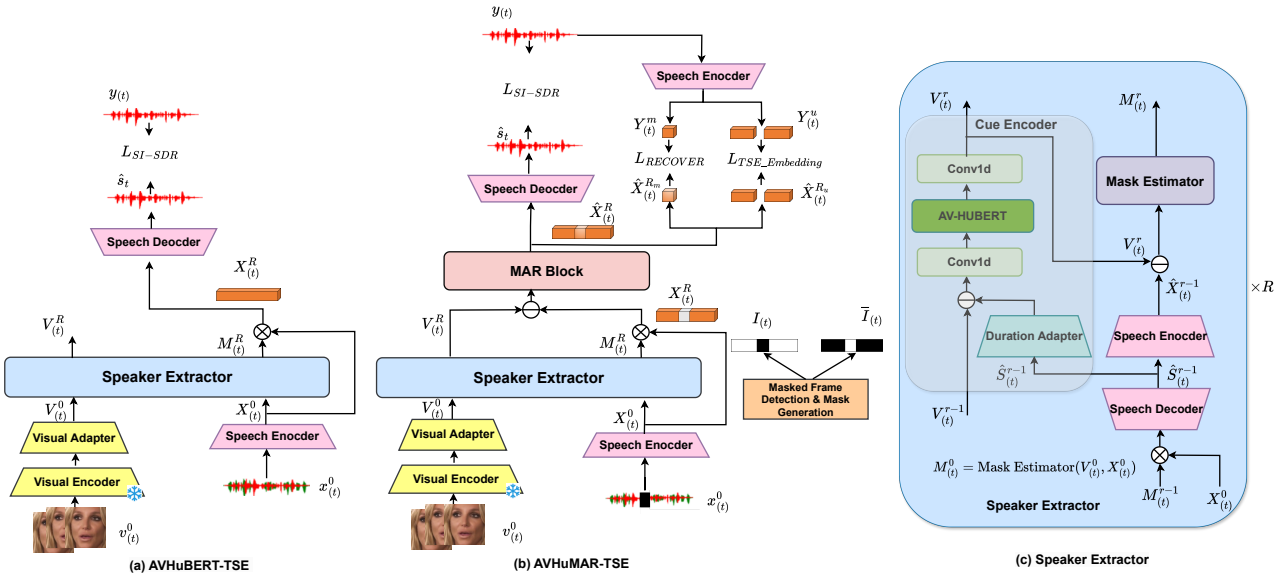

Figure 1. The overall architecture of the proposed AVHuMAR-TSE system

As shown in Fig 1 (b), an additional MAR Block will be inserted between the speech decoder and the final speaker extractor. As the name described, certain frames of the mixture speech waveform $x^0_{(t)}$ will be masked. Note that the masked region of the mixture speech embedding $X^0_{(t)}$ may be temporally compressed after passing through the speech encoder. For generalization consideration, an automatic masked frame detection step is necessary. According to the detected masked regions, a mask $I_{(t)}$ and an inverse mask $\overline{I}_{(t)}$ will be generated for loss computation.

The MAR Block plays two roles in the AVHuMAR-TSE system. First, it will be used to predict the masked target speech embedding. In this process, the unmasked target speech embedding region could provide rich speech context information and push the model to learn intra-modality correlation. The refined visual cues could push the model to learn a direct mapping between the target speaker's lip movements and the masked speech embedding region. Second, the MAR Block will be jointly optimized with the AVHuBERT-TSE system, thereby enabling further refinement of the estimated target speech embedding from the AVHuBERT-TSE system. To perform the embedding level loss, the ground truth of the target speech waveform $y_{(t)}$ will be input to a speech encoder to get its embedding $Y_{(t)}$.

2.3. Two-stage training strategy and loss functions

During the first training stage, only the scale-invariant signal-to-noise ratio (SI-SDR) will be calculated for the AVHuBERT-TSE system, which is shown in (1),

$$L_{SI-SDR}(y_t, \hat{s}_t) = -10\log_{10}\left(\frac{\frac{||<\hat{s}_t, y_t>y_t||^2}{||y_t||^2}}{||\hat{s}_t - \frac{<\hat{s}_t, y_t>y_t}{||y_t||^2}||}\right). \quad (1)$$

$$L_{RECOVER}(\hat{X}^{R_m}_{(t)}, Y^m_{(t)}) = MSE(\hat{X}^{R_m}_{(t)}, Y^m_{(t)}),$$
$$L_{TSE\_Embedding}(\hat{X}^{R_u}_{(t)}, Y^u_{(t)}) = MSE(\hat{X}^{R_u}_{(t)}, Y^u_{(t)}),$$

where,

$$\hat{X}^{R_m}_{(t)} = \hat{X}^R_{(t)} \odot \overline{I}_{(t)}, Y^m_{(t)} = Y_{(t)} \odot \overline{I}_{(t)},$$
$$\hat{X}^{R_u}_{(t)} = \hat{X}^R_{(t)} \odot I_{(t)}, Y^u_{(t)} = Y_{(t)} \odot I_{(t)}. \quad (2)$$

During the second training stage, besides SI-SDR, two extra mean squared error (MSE) [3] losses will be added. To recover the masked embedding region, one MSE loss will be added to the masked region of $\hat{X}^R_{(t)}$, which is called $L_{RECOVER}$. To refine the TSE performance, another MSE loss will be added to the unmasked region of $\hat{X}^R_{(t)}$, which is called $L_{TSE\_Embedding}$. The details are described in (2), Specifically, $\hat{X}^{R_m}_{(t)}$ and $Y^m_{(t)}$ denote the masked region of predicted target speech embedding and corresponding ground truth embedding. $\hat{X}^{R_m}_{(t)}$ and $Y^m_{(t)}$ are obtained by the element-wise multiplication of the inverse mask $\overline{I}_{(t)}$ and $\hat{X}^R_{(t)}$, $Y_{(t)}$, respectively. The $\hat{X}^{R_u}_{(t)}$ and $Y^u_{(t)}$ are obtained in the similar way.

$$\begin{aligned} &L(y_t, \hat{s}_t, \hat{X}^{R_m}_{(t)}, Y^m_{(t)}, \hat{X}^{R_u}_{(t)}, Y^u_{(t)}) \\ =&\alpha * L_{SI-SDR}(y_t, \hat{s}_t) + \beta * L_{RECOVER}(\hat{X}^{R_m}_{(t)}, Y^m_{(t)}) \\ &+ \gamma * L_{TSE\_Embedding}(\hat{X}^{R_u}_{(t)}, Y^u_{(t)}), \end{aligned} \quad (3)$$

The loss function for the AVHuMAR-TSE system is presented in (3), we use α, β, γ as scale factors to balance three parts, respectively.

3. Experiments

Dataset We simulate a 2-speaker mixture dataset from the VoxCeleb2 [2] dataset. The dataset splits are similar to [10].

Table 1. AVHuMAR-TSE and baseline performances on test set.

Model	Cue Type	SI-SDR(↑)	SI-SDRi(↑)	SDR(↑)	PESQ(↑)	STOI(↑)
AV-ConvTasNet[17]	Lip	10.725	10.771	11.099	2.592	0.859
USEV[9]	Lip	10.785	10.829	11.332	2.646	0.862
MuSE[10]	Lip + Speaker	11.458	11.506	11.836	2.706	0.873
AVHuBERT-TSE	Lip	11.728	11.771	12.043	2.765	0.878
AVHuMAR-TSE	Lip	**12.331**	**12.379**	**12.726**	**2.922**	**0.887**

Additionally, all the utterances are clipped to 4 seconds during training and 4-6 seconds during inference. For the second training stage, a random segment of each utterance in the training set will be masked with zero value. To find the optimal mask duration of the proposed AVHuMAR-TSE system, we simulate the mask duration gap equal to 100ms, 200ms, 300ms, 400ms, 500ms, and 600ms, respectively. Each mask duration gap is applied to the entire training set.

Baselines and evaluation metrics Since the AVHuMAR-TSE system is a time-domain AV-TSE system. To make a fair comparison, we also select three time-domain AV-TSE systems including AV-ConvTasNet [17], USEV [9], and MuSE [10] as our baseline systems. For the evaluation metrics, we select the SI-SDR [14], the SI-SDR improvement (SI-SDRi), and the signal-to-noise ratio (SDR) [14] as subjective metrics. We use the perceptual evaluation of speech quality (PESQ) [13] and the short-term objective intelligibility STOI [16] as objective metrics. The higher the better for all metrics.

Implementation details We re-implement three baseline systems with float32 precision. The speaker extractor block number R is 4. The scale factors (α, β, γ) are set to (1,5,1). For the first training stage, we train 150 epochs with a learning rate of 0.00015. For the second training stage, we load the best checkpoint from the first training stage and then train for another 30 epochs with the same learning rate. Both training stages are conducted on 3 32G V100 GPUs with a batch size of 2.

4. Experimental results

4.1. Comparison with different baseline systems

As shown in Table 1, Compared to the baseline results, our proposed AVHuMAR-TSE could achieve the best performance in terms of all the metrics. With 12.331 on SI-SDR, 12.726 on SDR, 2.922 on PESQ, and 0.887 on STOI. Such results are significantly higher than the baseline results, which demonstrate the effectiveness of the cue encoder and the proposed MAR strategy.

To explore the improvements brought by each module, we also report the performance of AVHuMAR-TSE without MAR blocks. As shown in Table 1, the system could still achieve 12.043 on SDR, 2.765 on PESQ, and 0.878 on STOI, respectively. Both subjective and objective performances are better than MuSE. Note that MuSE utilizes the same visual cue for all mask estimators and still needs additional speaker labels as input. It is worth noting that after utilizing the MAR strategy, the SI-SDR can be further improved from 11.728 to 12.331 while PESQ improved from 2.765 to 2.922, and STOI improved from 0.878 to 0.887.

Table 2. Effect of different mask duration on SI-SDR

Mask Duration(ms)	SI-SDR(↑)	SI-SDRi(↑)
100	12.292	12.338
200	11.956	12.012
300	**12.331**	**12.379**
400	11.925	11.973
500	11.826	11.873
600	11.695	11.742

4.2. Effect of different mask durations

To investigate the effect of various mask durations on the final target speech extraction performance for AVHuMAR, we report the SI-SDR and SI-SDRi with mask duration increasing from 100 ms to 600 ms, with 100 ms as the interval. As shown in Table 2, when the mask duration is set to 300 ms, AVHuMAR achieves the best performance in terms of SI-SDR and SI-SDRi. Furthermore, the performances with mask durations of 100 ms and 200 ms are slightly better than the performances with mask durations of 400 ms and above. Except for the mask duration equal to 600 ms, all the SI-SDR and SI-SDRi results are higher than the results without the MAR strategy.

With guidance from both intermediate estimated target speech context and target speaker lip movements, the boundary of the masked region could be relatively easy to recover. However, the center area of the masked region could be hard to recover even with guidance from both modalities. Based on this analysis, too much gap could not be conducive to the MAR strategy. On the contrary, it may even bring some adverse effects. The model may learn some corrupted audio-visual correlations and the overall model weights might be biased towards the recovery task and forget the TSE knowledge learned in the first training stage.

4.3. Case study on AVHuMAR-TSE results

As shown in Fig 2, we visualize two cases. For case 1, AVHuMAR-TSE could achieve 13.943 in terms of SI-SDR while MuSE could only achieve -0.105. We mark three obvious extraction failed regions with green boxes in

the second spectrogram from MuSE including one high-frequency part and two low-frequency parts. For case 2, MuSE achieves SI-SDR with 7.931 while the AVHuMAR-TSE system could achieve 12.993. This time MuSE extracts the coarse target speech but still lacks certain low-frequency components, as indicated in the left-bottom box. Additionally, MuSE mistakenly extracts some high-frequency parts from another speaker, as highlighted in the right-top box.

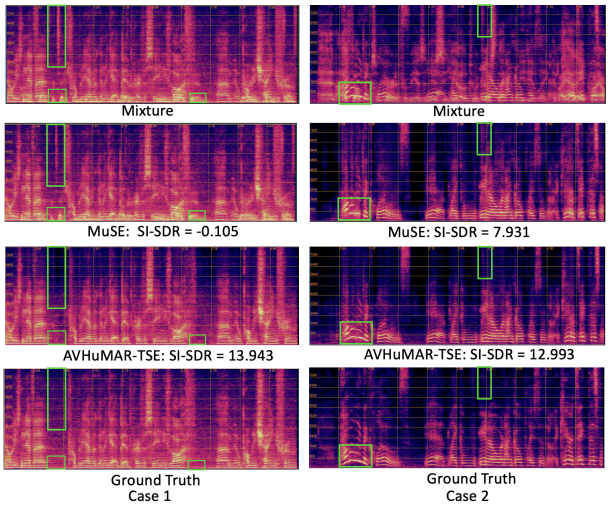

Figure 2. Comparison of target speech spectrograms extracted by AVHuMAR-TSE system and MuSE system.

5. Conclusion

In this study, we integrate pre-trained AV-HuBERT layers into the AV-TSE system as cue encoder and further propose the AVHuMAR-TSE system with MAR strategy. When compared to three time-domain AV-TSE systems, AVHuMAR shows substantial performance advancements in both subjective and objective metrics. The outcomes support the effectiveness of the cue encoder and the MAR strategy in enhancing audio-visual synchrony and speech context association. Through experiments, we find that an appropriate mask duration is crucial for the MAR strategy. In future work, we aim to further explore how the MAR strategy could align audio-visual latent feature space and expand our AVHuMAR-TSE system to various mixture scenarios.

References

[1] X. Chen, Y. Wang, X. Wu, D. Wang, Z. Wu, X. Liu, and H. Meng. Exploiting audio-visual features with pretrained av-hubert for multi-modal dysarthric speech reconstruction. In *IEEE ICASSP 2024*. 1

[2] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 2

[3] K. Das, J. Jiang, and J. Rao. Mean squared error of empirical predictor. 2004. 2

[4] R. Gao and K. Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. *2021 IEEE/CVF CVPR*, pages 15490–15500, 2021. 1

[5] M. Ge, C. Xu, L. Wang, C. E. Siong, J. Dang, and H. Li. Spex+: A complete time domain speaker extraction network. In *Interspeech*, 2020. 1

[6] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM TASLP*, 25(10):1901–1913, 2017. 1

[7] Y. Luo and N. Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM TASLP*, 27(8):1256–1266, 2019. 1

[8] S. Ouyang, R. Ye, and L. Li. WACO: Word-aligned contrastive learning for speech translation. In *ACL Long paper*, pages 3891–3907, July 2023. 1

[9] Z. Pan, M. Ge, and H. Li. Usev: Universal speaker extraction with visual cue. *IEEE/ACM TASLP.*, 30:3032–3045, sep 2022. 3

[10] Z. Pan, R. Tao, C. Xu, and H. Li. Muse: Multi-modal target speaker extraction with visual cues. In *ICASSP 2021*, pages 6678–6682, 2021. 1, 2, 3

[11] Z. Pan, R. Tao, C. Xu, and H. Li. Selective listening by synchronizing speech with lips. *IEEE/ACM TASLP*, 30:1–1, 01 2022. 1

[12] Z. Pan, W. Wang, M. Borsdorf, and H. Li. Imaginenet: Target speaker extraction with intermittent visual cue through embedding inpainting. In *ICASSP 2023*, pages 1–5, 2023. 1

[13] 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 *IEEE ICASSP*, volume 2, pages 749–752 vol.2, 2001. 3

[14] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey. Sdr – half-baked or well done? *IEEE ICASSP 2019,*, pages 626–630, 2018. 3

[15] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations*, 2022. 1

[16] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE ICASSP*. 3

[17] J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, and D. Yu. Time domain audio visual speech separation. In *2019 IEEE ASRU*, pages 667–673, 2019. 1, 3

[18] C. Xu, W. Rao, E. S. Chng, and H. Li. Spex: Multi-scale time domain speaker extraction network. *IEEE/ACM TASLP*, 28:1370–1384, 2020. 1

[19] X. Yue, J. Ao, X. Gao, and H. Li. Token2vec: A joint self-supervised pre-training framework using unpaired speech and text. In *ICASSP 2023*, pages 1–5, 2023. 1

[20] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černocký, and D. Yu. Neural target speech extraction: An overview. *IEEE Signal Processing Magazine*, 40(3):8–29, 2023. 1

[21] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černocký. Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):800–814, 2019. 1