Improving the Robustness of Audio-Visual Target Speaker Extraction With AV-HuBERT Based Lip Features

Jiarong Du[1,2], Zhan Jin[3], Bang Zeng[3], Peijun Yang[1,2], Ming Li[3], and Juan Liu[2,1] ⋆

[1] School of Cyber Science and Engineering, Wuhan University, Wuhan, China
[2] School of Artificial Intelligence, Wuhan University, Wuhan, China
[3] School of Computer Science, Wuhan University, Wuhan, China

Abstract. The target speaker extraction task aims to extract clean speech of the target person from a segment of mixed speech. In recent years, audio-visual speech enhancement (AVSE) has been increasingly applied, and the use of visual information from the target speaker has important application value in noisy environments. However, existing AVSE methods often face the problem of insufficient robustness of visual features, especially when parts of the content are missing or the video quality is poor. This will significantly reduce the effectiveness of the extracted visual features, further affecting the extraction performance. To address this issue, this paper first introduces a power compression strategy to enhance the effective components of the speech signal and avoid overreliance on visual information. Then, an end-to-end training approach is adopted to optimize the feature extraction process, initially alleviating the problem of insufficient robustness of lip movement features. To further improve performance, this paper uses the self-supervised AV-HuBERT model to extract features of lip movement.Through its multimodal self-supervised learning strategy, it can capture more discriminative dynamic features of lip movements and also achieve deep consistency between audio and video features. Experimental results show that the proposed method achieves stable improvements in key metrics such as PESQ, STOI, and SI-SDR, verifying the importance of visual feature extraction in the AVSE task and providing ideas for target speaker extraction in complex scenarios.

Keywords: audio-visual target speaker extraction · AV-HuBERT · speech enhancement

1 Introduction

In communication systems, the significance of Target Speaker Extraction (TSE) has become increasingly prominent. With the in-depth integration of voice interaction technologies into scenarios such as intelligent terminals, remote conferences, and autonomous driving, people's demands for voice communication

⋆ † Corresponding Author, E-mail: liujuan@whu.edu.cn

quality have been continuously rising. As a core means to address the issues, TSE has demonstrated increasingly notable value. TSE refers to the task of extracting clear speech from a target speaker within complex acoustic environments. It is analogous to the well-known classic "cocktail party problem"[2] — just as humans can accurately focus on a target conversationalist's voice at a noisy cocktail party, TSE needs to achieve a similar form of auditory focusing amid multiple interferences with the help of the target speaker information. Such environments typically contain adverse factors like background noise and overlapping speech from different speakers. The technical challenge lies in thoroughly eliminating various interferences while preserving as many details of the target voice as possible, ensuring that the extracted speech possesses high clarity and intelligibility.

In previous methods, researchers often utilized the target speaker's speech as registration information to assist the network in extracting the target speaker's voice from mixed speech. Traditional audio-only methods[28,32,7,33,9] rely solely on acoustic signals, completing the extraction task by analyzing speech features such as spectral characteristics and speaker information. Although they perform adequately in common environments, their effectiveness diminishes in complex scenarios with high noise and strong reverberation. In such cases, the acoustic features of the target speech are severely distorted, making accurate separation difficult using audio information alone. Furthermore, the target speaker's speech is not easily accessible in many situations. For instance, when dealing with unfamiliar speakers or temporary conversational scenarios, the approach of pre-collecting registered speech has obvious limitations.

In contrast, audio-visual methods[30] combine visual cues (e.g., lip movements) with audio data, leveraging multimodal synergy to enhance the robustness of speaker extraction. Visual information is unaffected by the acoustic environment, and lip movements have a natural synchronous relationship with speech content. This can provide additional identity information for the target speaker in complex environments, assisting in extracting the target speaker and compensating for the shortcomings of audio-only methods.

Despite significant progress in audio-visual speaker extraction methods in recent years[16], key challenges remain. It is widely acknowledged in research that the most speech-relevant information in the human face is lip movements. The deformation of the lips directly corresponds to articulatory actions, making it the most discriminative feature in the visual modality. However, how to correctly and appropriately encode lip movement information into usable features has long been a difficult problem. The network needs to capture both the instantaneous changes in lip shape and the continuous temporal characteristics. How to extract features to retain key information while removing redundancy is crucial for improving performance. Additionally, in many practical scenarios, there may be issues such as facial occlusion, video blurring, and reverberation. These problems can damage the integrity and accuracy of visual features, posing challenges to the extraction and utilization of visual features.

To address these issues, we have applied multiple modules. The most straightforward approach is to reduce audio-visual mismatch caused by reverberation. Reverberation can cause delay and distortion in audio signals, leading to a time difference between originally synchronized lip movements and speech. So we introduced the power compression method[11], which effectively suppresses the energy diffusion caused by reverberation through nonlinear compression processing of the audio signal's power spectrum. This makes the audio features more similar to the original vocalization state. It can improve the performance of the separation model in reverberant environments.

Subsequently, we focused on how to obtain more effective feature representations from limited visual information. To better match lip movement features with the backend separation network, we improved the traditional step-by-step training mode and performed joint training of the visual feature extractor and the separation model together. During the training process, the loss of the separation model is backpropagated to the visual feature extractor, guiding it to learn visual features that are more valuable for the separation task. This end-to-end optimization method effectively improves the quality of the extracted features.

In addition, a key issue in multimodal fusion is the consistency between modalities. We should ensure a high degree of synchronization between audio and visual features in terms of semantics and timing. The self-supervised AV-HuBERT[21] model is trained in an unsupervised manner on a large amount of audio-visual data, enabling it to learn the inherent synchronous relationship between audio and video. Introducing this model into the separation system allows the synchronous knowledge it has learned to guide the matching process between lip movement features and audio features. It can also enhance the correlation between the two at the feature level.

Our experimental results show that after enhancing audio and video features through the aforementioned methods, the separation system has achieved significant improvements in key metrics such as PESQ, STOI, and SISDR. Moreover, we have obtained excellent results on the test set of AVSEC4., verifying the effectiveness of the proposed methods in complex scenarios.

2 Related Work

2.1 AV-TSE

In recent years, there has been a proliferation of research on audio-visual target speaker extraction. Early research[8] has explored the impact of various modalities on the extraction of speakers. The authors not only investigated the registered speaker's voice and visual information but also examined the role of the speaker's azimuth angle. Through a series of comparative experiments, the authors demonstrated that in complex environments, video provides more information than audio. Zexu Pan's research[18,17] mainly focuses on how to deeply fuse the visual feature with mixed audio to better extract the target speaker's speech. Kai Li's research[12,13], starting from the real structure of the human

brain and neural responses, attempts to simulate the interaction patterns between neural signals in the human brain to construct neural networks, achieving remarkable results.

TF-GridNet[29] is one of the most advanced models in the field of speech separation. Zexu Pan took the lead in integrating the video modality into this model and proposed AV-GridNet[19]. Zhan Jin[10] improved AV-GridNet by combining audio and visual features in the channel dimension before entering the separation module. It simplifies the network while improving its performance.

2.2 Feature Extraction Methods

Scholars have conducted various studies on network structures with the aim of exploring how to make better use of the synchronous information between audio and video. The quality and robustness of visual features directly affect the subsequent feature fusion and speech separation processes. Early studies have shown that lipreading tasks can leverage networks to learn the correlation between lip movements and the expressed content, thereby converting lip movement information into sound-related features. Most studies thus utilize features extracted by the aforementioned methods to assist in target speaker extraction. From the initial application of 3D convolution [3,6,27,24] to the subsequent construction of 3D+2D convolutional neural network structures [25,20,14], existing lipreading technologies have become relatively mature, with the 3D CNN combined with 2D ResNet structure being widely adopted. A commonly used model is [15], which has been applied in many well-known AVSE tasks in recent years. However, in many cases, human faces are not as clear and stable as those in datasets, which limits the quality of features extracted by lipreading networks. Moreover, this approach lacks the capture of temporal consistency among multimodal features.

AV-HuBERT[21] constructs a stronger feature representation network through self-supervised learning, utilizing the strong correlation between audio and lip movement information. Moreover, random missing of modalities is introduced during the training process to enhance the robustness of feature extraction. This method has been widely used in AVSR[22,23] and proven to be effective, but its application in the speech separation is still limited. Study [4] demonstrates that partially or fully fine-tuning pre-trained AV-HuBERT can effectively improve performance in AVSE by leveraging generalizable multi-modal embeddings. And AVHuMAR-TSE[31] integrates AV-HuBERT layers with a Mask-And-Recover (MAR) strategy to exploit speech context and refine visual-audio correspondence, while contrastive losses and iterative cue refinement mechanisms are employed to strengthen the consistency between lip movements and speech semantics. In contrast, we adopted a straightforward and simple approach: directly replacing the original lip-reading model with the AV-HuBERT model, with the only difference being that the extracted features have a higher dimension.

3 Methods

The overall architecture of the network is illustrated in Figure 1. Similar to common AV-TSE structures, the network mainly consists of an audio encoder, a video encoder, a speaker extractor, and a decoder. The extractor of this system is similar to [29], which is not the focus of this paper. To achieve good results on datasets that are complex and close to real-world scenarios, we have adopted various methods. The first part will briefly introduce our explorations on feature enhancement; the second part will focus on analyzing the method of extracting lip movement features using the pre-trained AV-HuBERT.

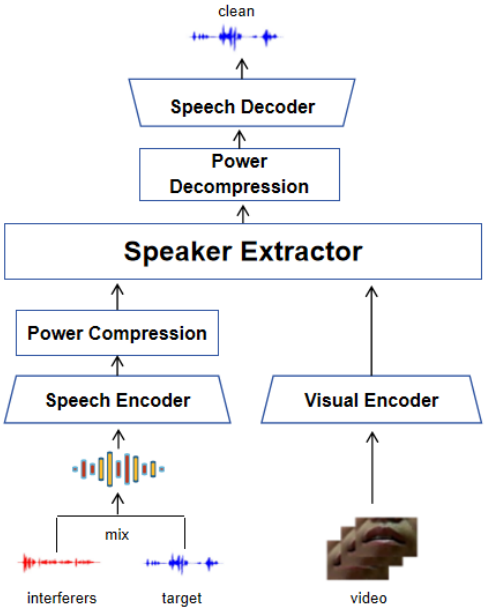


Fig. 1: The overall architecture of the proposed system.

3.1 Power Compression And End-to-End Training

In AVSE tasks, the effective utilization of video information is a key factor in achieving successful separation. Reverberation not only affects the acoustic structure of speech but also disrupts the synchronization between sound and lip movements. Therefore, the separation task becomes quite challenging when dealing with reverberant scenarios. According to [11], power compression and phase estimation methods can effectively handle acoustic scenarios in reverberant environments. Thus, we have integrated this method into our structure.

Specifically, for the input complex spectrogram representation $\mathbf{X} \in C^{B\times M\times T\times F}$ (where B denotes the batch size, M is the number of channels, T represents time frames, and F stands for frequency bins), we first decompose it into the magnitude spectrum $|\mathbf{X}|$ and phase spectrum $\angle\mathbf{X}$. A non-linear transform $||\mathbf{X}||_2^2$ is applied to the magnitude spectrum to compress high-power components, effectively mitigating the long reverberation tails and magnitude distortions caused by reverberation. Subsequently, we reconstruct the complex representation using the original phase information $\angle X$ and the compressed magnitude:

$$\mathrm{Re}(\mathbf{X}') = ||\mathbf{X}||_2^2 \cdot \cos(\angle\mathbf{X}), \quad \mathrm{Im}(\mathbf{X}') = ||\mathbf{X}||_2^2 \cdot \sin(\angle\mathbf{X}) \tag{1}$$

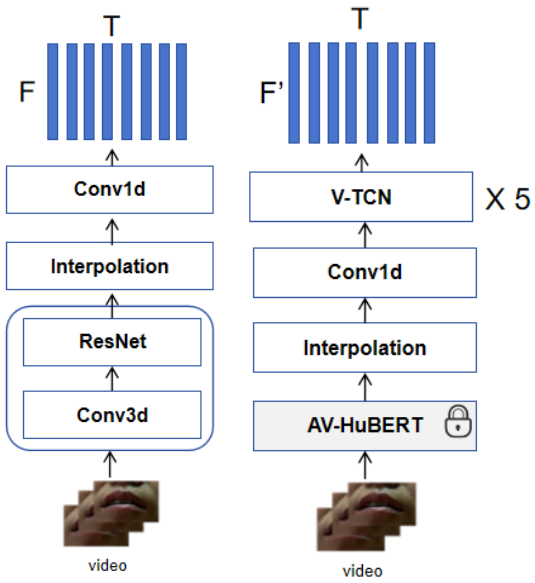
This processing ensures that phase information is fully retained, while the magnitude is optimally adjusted. Finally, the real part $\mathrm{Re}(X')$ and imaginary part $\mathrm{Im}(X')$ are concatenated along the channel dimension to form an extended feature $[\mathrm{Re}(X'), \mathrm{Im}(X')] \in R^{B\times 2M\times T\times F}$, which is fed into the subsequent network.

Most researchers utilize a pre-trained lip-reading network as the extractor for visual features. However, in complex scenarios, pre-trained networks may not be sufficient to fully represent a speaker's lip movement information as features. Therefore, to enhance the representation capability of visual features, we perform end-to-end training of the feature extraction network and the separation network. It enables the extracted features to better adapt to the subsequent separation process and improves the network's ability to extract the target speaker. As shown in Figure 2(a), we unlock the weights of the visual extractor and fine-tune them during the training process. And lip movement information is encoded into 512-dimensional features.

3.2 AV-HuBERT for Lip Feature Extraction

To further enhance the separation capability in complex acoustic environments, we focus on exploring how to use better features to represent the speaker's lip movement information. Studies in [34] and [26] have shown that in speaker extraction tasks, there may be cases of extracting incorrect targets, especially in complex scenarios. [34] also mentions that the reason for this result is likely that the features of the speaker's registration information are not strong enough. The network structure of commonly used methods [15] is relatively simple, and it lacks good robustness to special cases such as video frame loss, which may affect the subsequent speaker extraction results. Based on this, we propose to use the advanced large model AV-HuBERT to model the speaker's lip movement information.

Unlike the method in [31], we directly utilize the pre-trained Noise-Augmented AV-HuBERT Large weights to convert visual information into features. These weights were trained on two commonly used audio-visual datasets, LRS3[1] and VoxCeleb2[5], with the addition of noise to the training data. It can enhance the robustness in complex scenarios. Compared to ordinary lip-reading networks, AV-HuBERT is not only trained on a larger scale of data but also simulates modality missing during the training process, resulting in stronger feature extraction capabilities. Intuitively, using AV-HuBERT to extract video information



(a) Lipreading model (b) AV-HuBERT

Fig. 2: Detailed structure of the visual encoder.

can capture superior lip movement features, enabling the model to learn a better matching relationship between audio and video. The extracted features are 1024-dimensional, which doubles the information volume compared to previous ones. Therefore, we removed the original V-TCN module, reducing both the network size and the difficulty of learning. In the multi-layer separation modules, better lip movement features can learn more matching information from the mixed speech.

4 Experiments

4.1 Dataset

We used the provided training set for model training and the development set for validation in the 4th AVSE challenge. The training set contains 34,524 utterances (with a total duration of 113 hours and 17 minutes), involving 605 target speakers. The interferers are selected from a pool of 405 competing speakers and 7,346 noise files (covering 15 noise categories). The development set includes 3,306 utterances (8 hours and 38 minutes), involving 85 target speakers. The interferers are chosen from 30 competing speakers and 1,825 noise files (belonging to the same 15 noise categories as the training set).

The target data of the dataset is composed of videos from the LRS3 dataset, which includes spoken sentences extracted from TED and TEDx videos. The interferers are divided into three categories: speech interference, non-speech noise,

and music. Speech interference is also derived from the LRS3 dataset, with strict measures taken to ensure that the target speakers and interfering speakers belong to completely disjoint sets. Non-speech noise is sourced from two well-recognized datasets: the Clarity Enhancement Challenge (CEC1) and the Deep Noise Suppression (DNS) challenge. Music is extracted from the MedleydB multi-track music dataset, which contains 122 royalty-free songs.

The test set consists of 3,180 utterances, among which 1,500 are used for leaderboard evaluation, and the remaining are reserved for listening tests.In the set, the reverberation is not generated by simulation; instead, it is produced in real conference room scenarios by controlling the room size and sound propagation distance. And the signal-to-noise ratio (SNR) ranges from -18 dB to 6.55 dB.

4.2 Implementation details

Then we dynamically mixed the target and interferers at signal-to-noise ratios (SNRs) ranging from -18 dB to 6 dB during training. Mixed audio segments and corresponding target speaker videos (sampled at 25 frames per second) were randomly truncated into 3-second chunks. We employed the Adam optimizer with an initial learning rate of 0.001, and the learning rate was halved whenever the best validation loss showed no improvement over three consecutive epochs. The SI-SDR-SE loss, same as [10], was used as the training loss function, with the batch size fixed at 2.

4.3 Results and discussions

The detailed results are shown in Table 1 and 2. Based on the three objective metrics in the table, compared with the initial "Noisy" state, both the speech quality and intelligibility after separation are significantly improved. In comparison with the basic separation system, power compression (System I) exhibits obvious improvements in all three metrics. Similarly, end-to-end joint training (System II) further optimizes on the basis of System I and also provides a considerable degree of increment. Meanwhile, System III, which integrates the first two methods, achieves the optimal results. This fully verifies the effectiveness of the integration of multiple technologies in enhancing the quality and intelligibility of noisy speech, and provides a technical path reference for related tasks such as speech enhancement. Our System III ranks among the top two in all metrics and ranked first in the subjective evaluation.

Building on the above findings that validate the effectiveness of power compression and end-to-end training in enhancing separation performance, we further explored the potential of advanced visual feature extraction methods by introducing the AV-HuBERT model for lip movement feature extraction. To isolate the impact of AV-HuBERT on the system, we maintained the power compression strategy—consistent with the setup of System III—to ensure the audio preprocessing pipeline remained unchanged, while deliberately omitting end-to-end training. Instead, we directly adopted the lip movement features extracted

Table 1: Our results evaluated on 1,500 utterances in the test set

	separation	power compression	e2e	PESQ	STOI	SISDR
Noisy	×	×	×	1.285795	0.508202	-25.943447
baseline	√	×	×	1.875935	0.754096	-18.906067
System I	√	√	×	1.947982	0.772565	-18.297653
System II	√	×	√	2.069336	0.798168	-17.809357
System III	√	√	√	2.154918	0.824580	-17.138598

Table 2: All results evaluated on 1,500 utterances in the test set(WHU_DKU represents System III in Table 1.)

System	PESQ	STOI	SISDR
Noisy	1.285795	0.508202	-25.943447
Team-OPTIMAL	1.365186	0.479389	-21.439292
USTC_Entry1	1.350890	0.522072	-23.963609
GU-ENU	1.352315	0.523246	-23.925401
Rahma_Team	1.301840	0.545504	-24.722959
TeamKCW	1.578468	0.602192	-21.765331
SND_VD	1.720585	0.638246	-21.957590
R-test1	1.643556	0.657688	-19.258877
BioASP	1.714016	0.669129	-20.406594
SUSTechAILab	1.947145	0.671639	-18.744615
CITISIN	**2.290222**	0.779796	-17.144653
WHU_DKU(our)	2.154918	**0.824580**	**-17.138598**

by the pre-trained AV-HuBERT model as input to the separation network, aiming to assess the intrinsic superiority of these features in their raw form.

Notably, the initial leaderboard evaluation was conducted on a subset comprising approximately half of the total test set data, which may have limited the comprehensiveness of performance insights. To address this, we conducted a supplementary evaluation using the complete test set, ensuring a more robust and representative assessment of AV-HuBERT's capabilities across diverse acoustic scenarios, including varying levels of noise, reverberation, and speaker overlap.

As presented in Table 3, the results clearly demonstrate that even without the benefits of end-to-end fine-tuning, the features derived from AV-HuBERT still outperform those of our previously best performing System III across key metrics such as PESQ, STOI, and SI-SDR. This outcome not only underscores the robustness of AV-HuBERT's feature representation, which is shaped by self-supervised multimodal learning on a large scale, but also highlights its potential to reduce reliance on task-specific fine-tuning, offering a more generalizable solution for visual feature extraction in AVSE tasks.

Table 3: Evaluation results on the complete test set

System	PESQ	STOI	SISDR
Noisy	1.285795	0.508202	-25.943447
System III	1.467925	0.816043	-18.832981
AV-HuBERT	**1.534662**	**0.820336**	**-17.705436**

5 Conclusion

In this study, we explored the impact of various feature enhancement methods on the AVSE task and investigated the application of AV-HuBERT in lip movement feature extraction. Experimental results demonstrate that enhancing the representation capability of visual features can significantly improve the performance of the separation system. Using lip-reading models is not the only and best way to extract visual features; instead, adopting self-supervised models such as AV-HuBERT will deeply explore the connections between audio and video modalities and help the AVSE task achieve better outcomes.

References

1. Afouras, T., Chung, J.S., Zisserman, A.: Lrs3-ted: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496 (2018)
2. Arons, B.: A review of the cocktail party effect. Journal of the American Voice I/O society **12**(7), 35–50 (1992)

3. Assael, Y.M., Shillingford, B., Whiteson, S., De Freitas, N.: Lipnet: End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599 (2016)
4. Chern, I.C., Hung, K.H., Chen, Y.T., Hussain, T., Gogate, M., Hussain, A., Tsao, Y., Hou, J.C.: Audio-visual speech enhancement and separation by utilizing multi-modal self-supervised embeddings. In: 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). pp. 1–5. IEEE (2023)
5. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622 (2018)
6. Fung, I., Mak, B.: End-to-end low-resource lip-reading with maxout cnn and lstm. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2511–2515. IEEE (2018)
7. Ge, M., Xu, C., Wang, L., Chng, E.S., Dang, J., Li, H.: Spex+: A complete time domain speaker extraction network. arXiv preprint arXiv:2005.04686 (2020)
8. Gu, R., Zhang, S.X., Xu, Y., Chen, L., Zou, Y., Yu, D.: Multi-modal multi-channel target speech separation. IEEE Journal of Selected Topics in Signal Processing **14**(3), 530–541 (2020)
9. Hao, F., Li, X., Zheng, C.: X-tf-gridnet: A time–frequency domain target speaker extraction network with adaptive speaker embedding fusion. Information Fusion **112**, 102550 (2024)
10. Jin, Z., Zeng, B., Li, Z., Liu, X., Li, M.: A target speaker extraction method for the 3rd audio-visual speech enhancement challenge. System **1**(2.932), 0–876 (2024)
11. Li, A., Zheng, C., Peng, R., Li, X.: On the importance of power compression and phase estimation in monaural speech dereverberation. JASA express letters **1**(1) (2021)
12. Li, K., Xie, F., Chen, H., Yuan, K., Hu, X.: An audio-visual speech separation model inspired by cortico-thalamo-cortical circuits. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(10), 6637–6651 (2024)
13. Li, K., Yang, R., Sun, F., Hu, X.: Iianet: An intra-and inter-modality attention network for audio-visual speech separation. arXiv preprint arXiv:2308.08143 (2023)
14. Margam, D.K., Aralikatti, R., Sharma, T., Thanda, A., Roy, S., Venkatesan, S.M., et al.: Lipreading with 3d-2d-cnn blstm-hmm and word-ctc models. arXiv preprint arXiv:1906.12170 (2019)
15. Martinez, B., Ma, P., Petridis, S., Pantic, M.: Lipreading using temporal convolutional networks. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6319–6323. IEEE (2020)
16. Michelsanti, D., Tan, Z.H., Zhang, S.X., Xu, Y., Yu, M., Yu, D., Jensen, J.: An overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 1368–1396 (2021)
17. Pan, Z., Ge, M., Li, H.: Usev: Universal speaker extraction with visual cue. IEEE/ACM Transactions on Audio, Speech, and Language Processing **30**, 3032–3045 (2022)
18. Pan, Z., Tao, R., Xu, C., Li, H.: Muse: Multi-modal target speaker extraction with visual cues. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6678–6682. IEEE (2021)
19. Pan, Z., Wichern, G., Masuyama, Y., Germain, F.G., Khurana, S., Hori, C., Le Roux, J.: Scenario-aware audio-visual tf-gridnet for target speech extraction. In: 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 1–8. IEEE (2023)

20. Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., Pantic, M.: End-to-end audiovisual speech recognition. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 6548–6552. IEEE (2018)
21. Shi, B., Hsu, W.N., Lakhotia, K., Mohamed, A.: Learning audio-visual speech representation by masked multimodal cluster prediction. arXiv preprint arXiv:2201.02184 (2022)
22. Shi, B., Hsu, W.N., Mohamed, A.: Robust self-supervised audio-visual speech recognition. arXiv preprint arXiv:2201.01763 (2022)
23. Shi, B., Mohamed, A., Hsu, W.N.: Learning lip-based audio-visual speaker embeddings with av-hubert. arXiv preprint arXiv:2205.07180 (2022)
24. Son Chung, J., Senior, A., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6447–6456 (2017)
25. Stafylakis, T., Tzimiropoulos, G.: Combining residual networks with lstms for lipreading. arXiv preprint arXiv:1703.04105 (2017)
26. Tao, R., Qian, X., Jiang, Y., Li, J., Wang, J., Li, H.: Audio-visual target speaker extraction with selective auditory attention. IEEE Transactions on Audio, Speech and Language Processing (2025)
27. Torfi, A., Iranmanesh, S.M., Nasrabadi, N., Dawson, J.: 3d convolutional neural networks for cross audio-visual matching recognition. IEEE Access **5**, 22081–22091 (2017)
28. Wang, Q., Muckenhirn, H., Wilson, K., Sridhar, P., Wu, Z., Hershey, J., Saurous, R.A., Weiss, R.J., Jia, Y., Moreno, I.L.: Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking. arXiv preprint arXiv:1810.04826 (2018)
29. Wang, Z.Q., Cornell, S., Choi, S., Lee, Y., Kim, B.Y., Watanabe, S.: Tf-gridnet: Integrating full-and sub-band modeling for speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **31**, 3221–3236 (2023)
30. Wu, J., Xu, Y., Zhang, S.X., Chen, L.W., Yu, M., Xie, L., Yu, D.: Time domain audio visual speech separation. In: 2019 IEEE automatic speech recognition and understanding workshop (ASRU). pp. 667–673. IEEE (2019)
31. Wu, W., Chen, X., Wu, X., Li, H., Meng, H.: Target speech extraction with pre-trained av-hubert and mask-and-recover strategy. In: 2024 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2024)
32. Xu, C., Rao, W., Chng, E.S., Li, H.: Spex: Multi-scale time domain speaker extraction network. IEEE/ACM transactions on audio, speech, and language processing **28**, 1370–1384 (2020)
33. Zeng, B., Suo, H., Wan, Y., Li, M.: Sef-net: Speaker embedding free target speaker extraction network. In: Proc. Interspeech. pp. 3452–3456 (2023)
34. Zhao, Z., Yang, D., Gu, R., Zhang, H., Zou, Y.: Target confusion in end-to-end speaker extraction: Analysis and approaches. arXiv preprint arXiv:2204.01355 (2022)