Interspeech 2024
1-5 September 2024, Kos, Greece

Unified Audio Visual Cues for Target Speaker Extraction

Tianci Wu[1], *Shulin He*[1], *Jiahui Pan*[1], *Haifeng Huang*[2], *Zhijian Mo*[2], *Xueliang Zhang*[1]

[1]College of Computer Science, Inner Mongolia University, China
[2] Lenovo, Beijing, China

wutianci@mail.imu.edu.cn, cszxl@imu.edu.cn

Abstract

The target speaker extraction aims to isolate the target speaker's speech from other interfering speakers. Typically, an auxiliary reference, such as a pre-recorded speech or lip movements, is vital to direct attention to the target speaker. Existing methods use one of these cues or fuse both via attention mechanisms, yielding a shared feature of the target speaker. While both cues represent the same speaker, they have distinct attributes. The audio cue registers the speaker's timbre, but lip movements illustrate the synchrony. To blend the strengths of different cues, we propose a unified TSE network termed Uni-Net that employs a divide-and-conquer strategy to fuse audio and lip cues into distinct networks, capitalizing on each cue's unique information. Speech extracted from various cues acts as prior information, further refined by the post-processing network. We conducted the experiments on the public VoxCeleb2 corpus and Uni-Net achieves SOTA performance compared with baselines.
Index Terms: Target speaker extraction, audio-visual cues

1. Introduction

The objective of speech separation is to distinguish individual speakers' voices from a mixed signal, often referred to as the cocktail party problem [1]. It holds significant real-world applicability, encompassing fields such as speaker verification [2], speech recognition [3], and hearing aids [4]. Despite years of research, speech separation remains a non-trivial task. Recently, numerous methods have been proposed, including computational auditory scene analysis [5], non-negative matrix factorization [6], deep clustering [7]. However, these methods require prior information on the number of speakers, which is often impractical in real-world scenarios.

Human beings inherently possess the ability to focus on a target speaker while filtering out other interference [8]. Speaker extraction seeks to replicate this inherent human ability. It can isolate the speech signal of a target speaker from a mixed signal using auxiliary cues to identify the target. Speaker extraction can be categorized into two types based on the cues used: audio-only [8–10] and audio-visual [11–13] speaker extraction.

Audio-only speaker extraction employs a distinct speech sample of the target speaker as auxiliary information. The prerecorded utterance spoken by the target speaker is commonly encoded into a speaker embedding, which represents the speaker's timbre and vocal attributes. In [9], the extraction network utilizes the d-vector as the speaker embedding to estimate a spectrogram mask for the target speaker. In [8], the authors introduced a time-domain solution, wherein the speaker embedding network and the extraction network were trained jointly.

Although audio-only speaker extraction is universal, the performance is limited. Some factors, such as age, and emotions change the voice characteristics of the speaker. In addition, performance also degrades due to similar voice characteristics within the mixture. On the other hand, several studies [11,12,14] have started to use visual information, such as the face or lip image sequences. Different from audio cues, visual cues are temporally correlated with audio signals and unaffected by acoustic noise. Afouras et al. [15] use depth-wise convolution block to build an audio-visual network to predict a mask for the noisy magnitude spectrogram. Several studies [11, 12] also try to incorporate visual information into time-domain speaker extraction networks. While visual cues are invaluable in distinguishing the similar characteristics of different speakers, they can be affected by body movement and physical obstructions.

Different cues excel in various situations. Clearly, employing multiple cues concurrently can harness their strengths. Several studies have incorporated audio and visual cues into a shared network. In [16], audio and visual cues are initially added and subsequently fused with the mixed speech using Long Short-Term Memory (LSTM). In [17, 18], additive attention and normalized attention mechanisms are employed to integrate audio cues, visual cues, and mixed speech. Both audio and visual cues represent the same speaker, but they have distinct characteristics. Audio cues reflect the unique vocal attributes of a speaker, while visual cues are temporally synchronized with the target speech. Additionally, audio cues are time-invariant, while visual cues are time-variant, resulting in a discrepancy between the two. Consequently, integrating both audio and visual cues within a shared network with mixed speech not be the optimal choice. These cues will interfere with each other, significantly hinder the performance enhancement.

In this study, we introduce Uni-Net, a unified network for target speaker extraction that leverages both audio and visual cues. Our primary goal is to ensure that the two cues do not interfere with each other while simultaneously leveraging the strengths of both. Firstly, we employ a divide-and-conquer strategy to utilize individual TF-GridNet [19] networks to assimilate audio and visual cues, leveraging the distinct information provided by each cue. By adopting this strategy, we effectively eliminate conflicts between the cues, while simultaneously enhancing their respective strengths. Moreover, even if one cue is compromised by noise, the other remains capable of delivering speaker features without any mutual interference, thereby ensuring optimal extraction performance. By ensuring the target coherence of extracted speech derived from diverse cues and maintaining the alignment of time-frequency bins between the extracted speech and the mixture, we can decouple the speech of the target speaker from the mixture without requiring any auxiliary reference. To achieve this, we employ inplace convolution [20] as a post-processing network to analyze the interrelations among time-frequency bins and channels. On

10.21437/Interspeech.2024-934

the VoxCeleb2 dataset [21], experimental results demonstrate that our Uni-Net consistently surpasses other leading audio, visual, and audio-visual cues target speaker extraction models in both signal and perceptual quality.

2. Methodology

2.1. Overview

In this section, we describe the architecture of our unified target speaker extraction framework. Our proposed framework comprises three components: the audio-cues based extraction subnetwork (ACENet), the visual-cues based extraction subnetwork (VCENet), and the post-processing network (PPNet), as shown in Fig. 1. Broadly, audio and visual cues are processed through separate networks to harness their strengths and mitigate mismatches. Subsequent refinement via a post-processing network yields the target's speech.

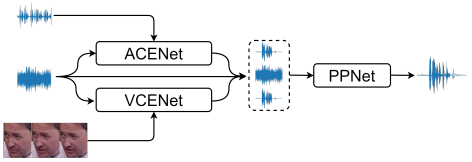


Figure 1: *Network architecture of our proposed Uni-Net.*

2.2. Speaker Extraction Network

We utilize TF-GridNet [19] to develop ACENet and VCENet, which represent state-of-the-art performance for speech separation. TF-GridNet operates within the STFT domain and adopts spectral mapping techniques to deduce the real and imaginary components of clean speech. The complex spectrum is encoded using a 2-dimensional convolution and a layer normalization. Every TF-Grid block comprises three main modules: the intra-frame spectral module, the sub-band temporal module, and the full-band self-attention module. The intra-frame spectral module interprets the input $R^{D\times T\times F}$ as T distinct sequences and deploys a BLSTM to capture full-band and spectral information for each frame. Analogously, the sub-band temporal module perceives the input $R^{D\times T\times F}$ as F individual sequences, leveraging a BLSTM to capture the temporal dynamics within each frequency. Within the full-band self-attention module, the input is reshaped into a representation of size $T \times (F \times D)$, where multi-head self-attention is employed to model global dependencies. In our work, the complex and magnitude spectrum of the mixture is fed into TF-GridNet.

To tailor TF-GridNet for multi-cues speaker extraction, we adopt a divide-and-conquer strategy that integrates audio and visual cues into distinct networks, as depicted in Fig. 2. We have designed specialized cue extractors for both audio and visual inputs, ensuring that each network is finely tuned to the nuances of its respective cues. This approach allows for a precise extraction by minimizing audio and visual cue interference.

2.2.1. Visual Cues Extractor

The visual cues extractor contains a 3-dimensional(3D) convolution layer, an 18-layer ResNet, and a visual temporal convolutional block (TCN) [12], as shown in Fig. 2, which seeks to extract the long-term visual representation from the lip image sequence. The 3D convolution captures spatio-temporal information from the image sequence. The 18-layer ResNet extracts higher-level visual features. The 3D convolutional layer and 18-layer ResNet are pre-trained from lip reading tasks as described in [22], with weights frozen during speaker extraction training. Each cropped lip image is encoded into a 512-dimentional representation by 3D convolutional layer and ResNet. Before feeding the feature into TCN, a linear layer is applied to compress 512 dimensions down to 128 to reduce complexity. The TCN contains a convolutional layer, a Relu activation, and a layer normalization, capturing the temporal content within a long-term lip spatio-temporal structure. The visual features are upsampled to match the audio feature sampling rate. We fuse the visual features and mixture with the convolutional layer.

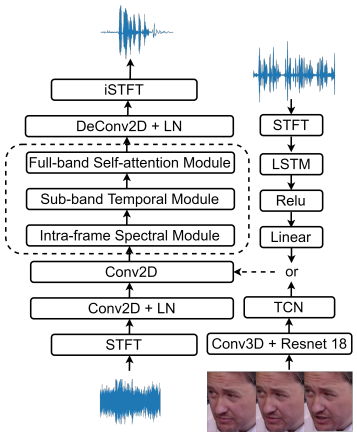


Figure 2: *Detailed structure of extraction network.*

2.2.2. Audio Cues Extractor

We derive audio cues via an audio cue extractor that incorporates a BLSTM layer, a ReLU activation function, and a linear layer. Initially, we compute the magnitude spectrum of the reference speech using an STFT with a 512-window size and a 256-hop size, corresponding to 32 ms and 16 ms at a 16 kHz sampling rate. The magnitude spectrum is then reshaped into a $(B \times T) \times F$ representation and further processed through a single-layer bidirectional LSTM with a hidden size of 32. A linear layer subsequently projects the BLSTM's output back to dimension 32. Furthermore, we apply mean-pooling to condense the time dimension of the reference representation.

By employing a divide-and-conquer strategy that uses separate networks to process audio and visual cues, we effectively and efficiently mitigate the mismatch between cues, while simultaneously enhancing their respective strengths. Even if one cue is compromised by noise, the other remains capable of delivering speaker representation without any mutual interference.

2.3. Post-Processing Network

Despite the distinct characteristics of audio and visual cues, the extracted speech uniformly converges towards the same target speech. Furthermore, the time-frequency correspondence between the extracted speech and the mixture provides vital prior information that can enhance the extraction of the target speech and suppress interference. With sufficient prior information at hand, we can extract the speech of the target speaker from the mixture without requiring any auxiliary reference.

We implement inplace convolution [20] as our post-processing network to analyze the relationship between preliminary extracted speech and mixture, as depicted in Fig. 3. The architecture comprises three main components: an Inplace Encoder, a Frequency-wise LSTM, and an Inplace Decoder. Each, the Inplace Encoder and Decoder, employ six layers of inplace

convolutional operations. In contrast to traditional convolution which typically downsamples features along the frequency dimension, inplace convolution adopts a stride of one, thereby preserving spectral details and facilitating the analysis of inter-channel correlations. To model the temporal dependencies, a two-layer BLSTM is employed on each frequency bin, processing only channel-wise features. This method facilitates independent frequency analysis while concurrently integrating information across various channels.

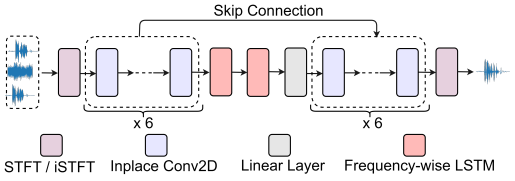


Figure 3: *Architecture of Post-Processing network.*

2.4. Loss Function

In our experiments, we use scale-invariant source-to-noise ratio (SI-SNR) [23] as loss function, which can be formulated as:

$$\begin{cases} s_{\text{target}} := \frac{\langle \hat{s}, s \rangle s}{\|\hat{s}\|^2} \\ e_{\text{noise}} := \hat{s} - s_{\text{target}} \\ \text{SI-SNR} := 10 \log_{10} \left(\frac{\|s_{\text{target}}\|^2}{\|e_{\text{noise}}\|^2} \right) \end{cases} \tag{1}$$

where $\hat{s} \in R^{1 \times T}$ and $s \in R^{1 \times T}$ are the extracted speech and target speech, respectively. $<\hat{s}, s>$ denote the signal power.

3. Experiments

3.1. Datasets

We evaluate the performance of our proposed framework using the VoxCeleb2 dataset [21]. From this dataset, we selected 48,000 utterances encompassing 800 speakers for training, and 36,237 utterances from 118 distinct speakers for testing, ensuring no speaker overlap between the two sets. We subsequently constructed a 2-speaker mixture dataset with 19,487, 4,766, and 2,867 utterances for the training, validation, and testing sets, respectively. During this process, interference speech was amalgamated with the target speech, using a randomized Signal-to-Noise (SNR) ratio that varied between -10dB and 10 dB. For visual cues, we utilized the video data corresponding to the target speaker present in the mixture. For the audio cues, we selected a random utterance from the same speaker that was not used in creating the speech mixture. The video is sampled at 25fps. The audio is resampled at 16kHz.

We simulate the visual occlusion by obscuring the visual signal from a random starting frame for an arbitrary duration. This replicates situations where the face-tracking method cannot identify the target speaker because of reasons including occlusion or body movement. We keep the audio cues intact.

3.2. Implementation Setup

We implemented the Uni-Net in accordance with the specifications delineated in Section 2. The same configuration parameters were applied to both ACENet and VCENet. The system utilizes a window and hop size of 32ms and 16ms, respectively, corresponding to a sampling rate of 16kHz. For the configuration of the GridNet block, the following parameters were established: D = 32, B = 3, I = 4, J = 1, H = 128, and L = 4. For the PPNet, the window and hop sizes were configured to 20ms and 10ms, respectively. The outputs from ACENet, VCENet, and the original mixture were concatenated along the channel dimension to facilitate the generation of real and imaginary components through STFT. The number of channels in PPNet is 48.

The models were trained with an Adam optimizer [24] with a learning rate of 0.001. If the validation loss does not decrease for 3 epochs, the learning rate is halved. Training is terminated if the loss does not decrease for 5 epochs.

3.3. Evaluate Metrics

The performance is evaluated with four common objective metrics: perceptual evaluation of speech quality (PESQ) [25], short-time objective intelligibility (STOI) [26], Signal-to-distortion ratio (SDR) [27], and Scale-invariant source-to-noise ratio (SI-SNR) [23]. STOI values typically range from 0 to 1, which can be roughly viewed as percent correct. PESQ values range from -0.5 to 4.5, which measures the overall perceptual quality. SI-SNRi measures the speech signal quality. For both metrics, the higher number indicates better performance.

3.4. Baselines

For the speaker extraction using visual cues, TDSE [13], MuSE [11], and AVDPRNN [12] serve as our baseline models. We retain the optimal configurations for TDSE and MuSE as established in their respective publications. For AVDPRNN, which stands as one of the top-performing networks for audio-visual speech extraction, we have increased the number of parameters and computational complexity to further improve its performance. For speaker extraction based on audio cues, we have compared our framework with SpEx+ [8] and have also developed a DPRNN system by integrating a speaker encoder with a configuration analogous to that of SpEx+. Our multi-cues speaker extraction framework has been subjected to a comparative analysis with SpeakerBeam [18], which employs an attention mechanism to integrate audio and visual cues. Furthermore, we have altered the SpeakerBeam framework by substituting the attention mechanism with a concatenation operator, hereby referenced as SpeakerBeam*.

To assess our proposed framework's efficacy in scenarios where one cue is obstructed, we conducted comparative evaluations with TDSE and ImagineNet [28]. TDSE lacks visual refinement modules, whereas ImagineNet exemplifies state-of-the-art (SOTA) performance in audio-visual speaker extraction when visual cues are occluded.

4. Results

4.1. Comparison With Baseline

Table 1 displays the metric scores for speaker extraction systems utilizing audio cues, visual cues, and a combination of audio-visual cues. When comparing SpeakerBeam, which utilizes normalized attention to fuse audio and visual cues, with TDSE, MuSE, ConvTasNet, and SpEx+, which rely on either audio or visual cues, it is evident that the integration of multi-cues generates a more accurate representation of the target speaker than single-cue approaches, hereby improving speaker extraction performance. In addition, SpeakerBeam outperforms SpeakerBeam* which can be attributed to the attention mechanism's ability to dynamically weigh audio and visual cues.

Our proposed framework significantly outperforms SpeakerBeam, achieving improvements of 1.72 dB in SI-

Table 1: *A comparison across various speaker extraction implementations on VoxCeleb2 2-speaker mixture, using audio, visual, or audio-visual as auxiliary cues. The SI-SNR, SDR, PESQ, and STOI of the mixture are -0.09, 0.00, 1.24, and 0.63, respectively.*

Cues Type	Methods	SI-SNRi (-0.09)	SDRi (0.00)	PESQ (1.24)	STOI (0.63)	Parameters
Visual	TDSE	11.10	11.44	2.16	0.862	13.70M
	MuSE	11.59	11.91	2.20	0.872	15.00M
	AVDPRNN	12.62	12.94	2.30	0.888	11.88M
Audio	DPRNN	11.76	12.07	2.21	0.868	10.95M
	SpEx+	10.37	10.68	2.08	0.850	11.50M
Audio-Visual	SpeakerBeam	11.91	12.17	2.20	0.873	11.38M
	SpeakerBeam*	11.30	11.65	2.16	0.864	11.38M
	Uni-Net	**13.63**	**13.80**	**2.66**	**0.906**	**5.01M**

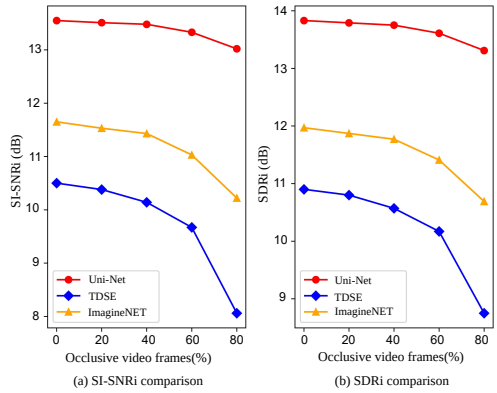
Table 2: *Ablation studies for Uni-Net. We study the contribution of different components. A, V, and P represent audio cues, visual cues, and post-processing modules, respectively.*

Methods	Audio	Visual	PPNet	SI-SNRi	PESQ
Uni-Net	✓	✓	✓	13.63	2.66
-w/o V	✓	×	✓	13.38	2.63
-w/o A	×	✓	✓	13.28	2.58
-w/o V, P	✓	×	×	13.07	2.63
-w/o A, P	×	✓	×	12.74	2.54

Table 3: *Performance of Uni-Net with baselines under visual availability or missing during training and inference phases.*

Methods	V Missing	SI-SNRi	PESQ	SDRi
ImagineNET	×	12.51	2.30	12.81
TDSE	×	11.31	2.17	11.63
Uni-Net	×	**13.63**	**2.66**	**13.80**
ImagineNET	✓	10.92	2.16	11.31
TDSE	✓	9.51	2.04	10.02
Uni-Net	✓	**13.30**	**2.62**	**13.58**

SNRi, 1.63 dB in SDRi, 0.46 in PESQ, and 3.3% in STOI, respectively. It can be attributed to the fact that, instead of integrating audio and visual cues into a shared network, our framework uses a divide-and-conquer strategy to integrate them into different networks, allowing for more tailored and effective cue-specific speaker extraction. In addition, by using a post-processing network, the relationship between time-frequency bins and channels can be effectively modeled, thereby further enhancing performance.



(a) SI-SNRi comparison (b) SDRi comparison

Figure 4: *SI-SNRi (a) and SDRi (b) score under different occlusion ratios of the video frame.*

4.2. Ablation Study

To further demonstrate the validity of our proposed model, we conducted an ablation analysis and designed four variants of Uni-Net. Table 2 shows the performance of variants.

The results indicate that excluding visual cues leads to a reduction of 0.25 dB in SI-SNRi and 0.03 in PESQ scores. The absence of audio cues results in a more serious decrease of 0.35 dB in SI-SNRi and 0.8 in PESQ. The absence of either audio or visual cues, when combined with the exclusion of the post-processing network, results in a significant degradation. The results demonstrate that incorporating both audio and visual cues substantially enhances the efficacy of speaker extraction, while the inclusion of a post-processing network further mitigates interference, capitalizing on the preliminary extracted speech.

4.3. Occlusion Study

We further assess the robustness of Uni-Net and baselines when visual cues are occluded. As indicated in the table 3, TDSE exhibits the lowest performance, which is attributable to the absence of visual refinement modules, whereas ImagineNET surpasses TDSE by using an audio-visual joint representation to refine occluded visual cues. The improvement demonstrated by Uni-Net can be attributed to the effectiveness of audio cues; even in the absence of visual cues, audio cues can still provide a precise representation of the speaker.

We have graphed the improvements in SI-SNRi and SDRi across varying percentages of occluded video frames. We determined varying mask lengths by considering the percentage of occlusion relative to the total video duration. In Figure 4, Uni-Net outperforms the other baselines and achieves consistent improvement even when 80% of the video is occluded.

5. Conclusion

In this paper, we propose a unified target speaker extraction framework to overcome the conflict between audio and visual cues. Specifically, we utilize a divide-and-conquer approach, integrating audio and visual cues into distinct subnetworks to capitalize on the unique information each cue offers. Moreover, this strategy ensures consistent performance enhancement, even when visual cues are obscured. We introduce a Post-Processing network to further extract target speech by analyzing the relationship between time-frequency bins and channels based on the initial extracted speech from various cues and mixtures. Our proposed network has been compared with other leading speaker extraction methods that employ audio, visual, and audio-visual cues. The results indicate that it achieves superior performance in comparison to other competitive baselines.

Acknowledgments: This research was partly supported by the China National Nature Science Foundation (No. 61876214).

6. References

[1] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[2] B. Liu and Y. Qian, "ECAPA++: Fine-grained Deep Embedding Learning for TDNN Based Speaker Verification," in *Proc. INTERSPEECH 2023*, 2023, pp. 3132–3136.

[3] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 9, no. 5, pp. 1–28, 2018.

[4] D. Wang, "Deep learning reinvents the hearing aid," *IEEE spectrum*, vol. 54, no. 3, pp. 32–37, 2017.

[5] G. Hu and D. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 2, pp. 396–405, 2007.

[6] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[7] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[8] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Spex+: A complete time domain speaker extraction network," in *INTERSPEECH*, 2020, pp. 1406–1410.

[9] H. R. Muckenhirn, I. L. Moreno, J. Hershey, K. Wilson, P. Sridhar, Q. Wang, R. A. Saurous, R. Weiss, Y. Jia, and Z. Wu, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *conference of the international speech communication association*, 2019.

[10] S. He, H. Li, and X. Zhang, "Speakerfilter: Deep learning-based target speaker extraction using anchor speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 376–380.

[11] Z. Pan, R. Tao, C. Xu, and H. Li, "Muse: Multi-modal target speaker extraction with visual cues," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6678–6682.

[12] Z. Pan, M. Ge, and H. Li, "Usev: Universal speaker extraction with visual cue," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 3032–3045, 2022.

[13] 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 automatic speech recognition and understanding workshop (ASRU)*. IEEE, 2019, pp. 667–673.

[14] J. Lin, X. Cai, H. Dinkel, J. Chen, Z. Yan, Y. Wang, J. Zhang, Z. Wu, Y. Wang, and H. Meng, "Av-sepformer: Cross-attention sepformer for audio-visual target speaker extraction," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[15] T. Alfouras, J. Chung, and A. Zisserman, "The conversation: deep audio-visual speech enhancement," in *Interspeech*, vol. 2018. International Speech Communication Association, 2018.

[16] T. Afouras, J. S. Chung, and A. Zisserman, "My lips are concealed: Audio-visual speech enhancement through obstructions," *Interspeech 2019*, 2019.

[17] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, "Multimodal speakerbeam: Single channel target speech extraction with audio-visual speaker clues." in *INTERSPEECH*, 2019, pp. 2718–2722.

[18] H. Sato, T. Ochiai, K. Kinoshita, M. Delcroix, T. Nakatani, and S. Araki, "Multimodal attention fusion for target speaker extraction," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 778–784.

[19] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "Tf-gridnet: Making time-frequency domain models great again for monaural speaker separation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[20] J. Liu and X. Zhang, "Inplace Gated Convolutional Recurrent Neural Network for Dual-Channel Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 1852–1856.

[21] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[22] T. Stafylakis and G. Tzimiropoulos, "Combining Residual Networks with LSTMs for Lipreading," in *Proc. Interspeech 2017*, 2017, pp. 3652–3656.

[23] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[25] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality (PESQ) the new itu standard for end-to-end speech quality assessment part i–time-delay compensation," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 755–764, 2002.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "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, 2011.

[27] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[28] Z. Pan, W. Wang, M. Borsdorf, and H. Li, "Imaginenet: Target speaker extraction with intermittent visual cue through embedding inpainting," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.