Plug-and-Play Co-Occurring Face Attention for Robust Audio-Visual Speaker Extraction

Zexu Pan[1], Shengkui Zhao[1], Tingting Wang[2], Kun Zhou[1], Yukun Ma[1], Chong Zhang[1], Bin Ma[1]

[1]Tongyi Lab, Alibaba Group, Singapore
[2]Nanjing University of Posts and Telecommunications, NanJing, China

`zexu.pan@alibaba-inc.com`

Abstract

Audio-visual speaker extraction isolates a target speaker's speech from a mixture speech signal conditioned on a visual cue, typically using the target speaker's face recording. However, in real-world scenarios, other co-occurring faces are often present on-screen, providing valuable speaker activity cues in the scene. In this work, we introduce a plug-and-play inter-speaker attention module to process these flexible numbers of co-occurring faces, allowing for more accurate speaker extraction in complex multi-person environments. We integrate our module into two prominent models: the AV-DPRNN and the state-of-the-art AV-TFGridNet. Extensive experiments on diverse datasets, including the highly overlapped VoxCeleb2 and sparsely overlapped MISP, demonstrate that our approach consistently outperforms baselines. Furthermore, cross-dataset evaluations on LRS2 and LRS3 confirm the robustness and generalizability of our method.

Index Terms: audio-visual, speaker extraction, speech separation, co-occurring faces, attention

1. Introduction

In the real world, speech signals are often contaminated by overlapping background speech. This poses significant challenges for speech-based applications, including automatic speech recognition and speaker verification in human-computer interaction [1]. Moreover, the presence of such noise degrades the quality of speech data, rendering its suitability for training large artificial intelligence models for speech understanding or generation tasks [2]. Consequently, there is a critical need to extract clean speech from mixed signals, a challenge commonly known as the "cocktail party problem" [3].

Speech separation (SS) is a widely studied approach to separate overlapping speech signals into individual tracks [4–10]. However, SS often lacks speaker identity association. In addition, SS typically requires prior knowledge of the number of speakers for unified processing, which limits its capacity as the number of speakers increases. To address this, speaker extraction methods have been developed [11–14], inspired by human auditory attention mechanisms. These methods directly extract the target speaker's voice by conditioning on a reference signal, such as a reference speech signal [15, 16], visual recording [17–20], or even brain activity [21–23]. Unlike speech separation, speaker extraction processes each speaker independently, avoiding identity and capacity limitations.

Among the auxiliary conditioning signals used in speaker extraction, visual cues have gained significant attention due to their robustness to acoustic noise. Typically, the synchronized face recordings extracted through face detection and tracking algorithms are used. Existing approaches often rely exclusively
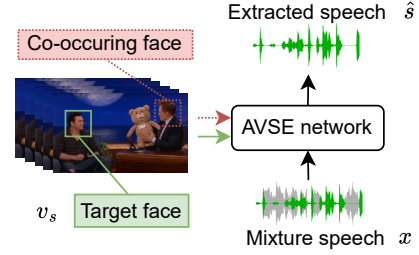

Fig. 1: *We explore the complementary speech activity cue offered by co-occurring on-screen face (red-dotted face box) when extracting the target (green non-dotted face box) speech in an audio-visual speaker extraction (AVSE) network.*

on the target speaker's visual cues, which the ambiguous visual-to-audio mappings (visemes to phonemes) frequently result in speaker confusion, over-suppression, and under-suppression errors in the extracted speech signals [24–27]. In real-world scenarios such as meetings, interviews, or social gatherings, multiple faces are often present on-screen. These co-occurring faces provide valuable speech activity cues that can enhance the accuracy of target speaker extraction. For instance, if a co-occurring speaker is inactive, there is a higher likelihood that a speech signal belongs to the target speaker, and vice versa, particularly in sparsely overlapped multi-person interactions. By modeling cross-speaker information in such complex environments, these errors can be significantly reduced.

In this paper, we address this gap by proposing a lightweight, plug-and-play inter-speaker attention module designed to leverage the complementary speech activity cues provided by co-occurring faces. The attention mechanism is capable of handling a flexible number of on-screen faces. Our proposed module could be seamlessly integrated into various existing audio-visual speech extraction (AVSE) architectures and iteratively explores inter-speaker correlations within the mixed speech signals during the extraction process.

To evaluate our approach, we integrate the proposed attention module into two representative AVSE models: (1) AV-DPRNN [28], a popular and prominent dual-path time-domain masking-based model, and (2) AV-TFGridNet [29], a state-of-the-art framework that leverages time-frequency domain modeling for direct-signal estimation. We conduct extensive experiments on diverse datasets, including VoxCeleb2 (highly overlapped speech) and MISP (sparsely overlapped speech) for 2-speaker and 3-speaker scenarios, demonstrating that our method consistently outperforms baseline models with only 0.2 million additional parameters, regardless of the number of co-occurring faces available. Furthermore, cross-dataset evaluations on LRS2 and LRS3 confirm the robustness and generalizability of our approach, underscoring its practical applicability.

2. Methodology

2.1. Fundamentals

Denote x as a multi-talker speech signal, which is the sum of the target speech signal s and interfering speech signals b_i:

$$x = s + \sum_{i=1}^{I} b_i \qquad (1)$$

where $i \in \{1, ..., I\}$ denotes the index of interfering speakers, the audio-visual speaker extraction network extracts an estimate of s denoted $\hat{s}$, from x, conditioned on the target speaker face recording v_s. Oftentimes, some interfering speakers' faces v_{b_i} co-occur on the screen along with the target speaker, here we explore such complementary speech activity information.

The majority of speaker extraction networks are inspired by speech separation architectures, typically comprised of a speech encoder, a speaker extractor, a speech decoder, and a visual encoder. The speaker extractor consists of multiple (R) repeated processing blocks, such as stacks of Temporal Convolutional Networks (TCNs) in Conv-TasNet [6, 18], Dual-Path networks in DPRNN [27, 30] or Sepformer [31, 32], or frequency-domain attention in TF-GridNet [5, 29]. In audio-visual speaker extraction, the output of the visual encoder is typically fused with one or all of these repeated blocks, enabling the network to leverage visual cues for improved speaker extraction performance.

2.2. Plug-and-play inter-speaker attention module

We introduce an Inter-Speaker Attention Module (ISAM) at the end of each repeated block in the speaker extractor, designed to be compatible with most speaker extraction networks. As illustrated in Figure 2, all available on-screen faces v_s and v_{b_i}, are fed into the AVSE network (merged along the batch dimension), and the network is conditioned to extract the signal associated with each face independently through the visual encoder and extractor stacks. At the end of each extractor stack, we incorporate the ISAM, which enables the target speech embedding to query all speaker's embeddings, and vice versa. The ISAM consists of a single self-attention layer along the speaker axis, followed by a layer normalization. We employ self-attention rather than cross-attention because the embedding representations of co-occurring faces share the same distribution as the target face, as essentially the co-occurring faces could also be potential target faces in other words. The extractor stack and ISAM are repeated R times to refine the extraction process.

To ensure the network remains robust in scenarios with no co-occurring faces, we occasionally bypass the ISAM during training. For scenarios involving an incomplete number of co-occurring faces, we employ random dropout of the co-occurring faces. This strategy enables the network to learn effective extraction even when co-occurring faces are absent or only partially available. Additionally, the ISAM is inherently robust to irrelevant on-screen faces (equivalent to silent faces in sparsely overlapping scenarios). This flexibility ensures that the module performs across a wide range of real-world conditions.

We integrate the ISAM into two popular AVSE models: AV-DPRNN [28] and AV-TFGridNet [29]. AV-DPRNN, known for its balanced performance in speech extraction and computational complexity, has demonstrated strong results on both highly overlapping [27] and sparsely overlapping speech [28]. In this model, we apply the ISAM to the embeddings after each stack of intra- and inter-chunk RNN processing. AV-TFGridNet, the current state-of-the-art AVSE model, operates
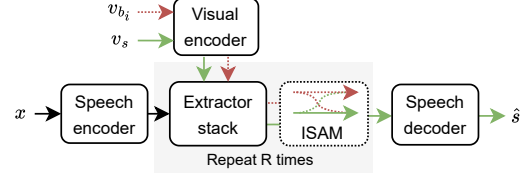

Fig. 2: *We introduce an optional (dotted box) inter-speaker attention module (ISAM) to compute attention to the co-occurring (dotted line) face activities. The plug-and-play ISAM is easily adaptable to the majority of AVSE networks.*

in the frequency domain. Here, we apply the ISAM to every time-frequency bin across speakers after each stack of intra-frame, sub-band, and full-band processing.

2.3. Objective function

We use the negative scale-invariant signal-to-noise ratio (SI-SNR) [33] as the objective function for training:

$$\mathcal{L}_{\text{SI-SNR}}(s, \hat{s}) = -20 \log_{10} \frac{\left| \frac{<\hat{s},s>}{|s|^2} s \right|}{\left| \hat{s} - \frac{<\hat{s},s>}{|s|^2} s \right|} \qquad (2)$$

The loss is applied to both the target speaker (s, $\hat{s}$), and the co-occuring speakers (b_i, $\hat{b}_i$).

3. Experimental setup

3.1. Datasets

We evaluate our proposed method and baselines using several audio-visual datasets: VoxCeleb2 [34], LRS2 [35], LRS3 [36], and MISP [37]. The VoxCeleb2 dataset is a multilingual, in-the-wild collection featuring diverse speakers. The LRS2 dataset comprises English speakers from BBC videos. The LRS3 dataset includes English speakers from TED videos, offering cleaner speech signals compared to VoxCeleb2 and LRS2. Lastly, the MISP dataset consists of Chinese speakers recorded in home conversation scenarios. These datasets collectively provide a comprehensive evaluation across varied languages, environments, and speech conditions.

The VoxCeleb2, LRS2, and LRS3 datasets consist exclusively of active speakers, making them suitable for simulating highly overlapped multi-talker speech. Following [38], we mix the target speech with interference speech at a Signal-to-Noise Ratio (SNR) randomly set between 10 dB and −10 dB. To maximize overlap, the longer speech clip is truncated to match the length of the shorter one. In contrast, the MISP dataset includes both active and inactive speakers, making it ideal for simulating sparsely overlapped multi-talker speech. Following [28], we select non-overlapping single-speaker segments as sources for simulation, mixing the interference speech with the target speech at an SNR randomly set between 10 dB and −10 dB. For this study, we simulate both 2-speaker and 3-speaker scenarios. All audio signals are sampled at 16 kHz, while video signals are sampled at 25 frames per second.

3.2. Training

For model implementation and experiments, we utilize the ClearerVoice-Studio toolkit [1]. During training, we employ the

---

[1] https://github.com/modelscope/ClearerVoice-Studio/

Table 1: *Results on VoxCeleb2 2-speaker mixture dataset. v-spks denotes the number of face tracks observed by the model: 1-spk for the target speaker only, 2-spk for both speakers.*

Sys.	Model	v-spks	SI-SNRi	SNRi	PESQi	STOIi
1	AV-DPRNN [28]	1-spk	11.5	11.8	0.89	0.24
2	AV-DPRNN-SS	2-spk	10.7	11.0	0.77	0.22
3	AV-DPRNN-SE	2-spk	11.2	11.5	0.83	0.23
4	AV-DPRNN-ISAM	1-spk	11.7	12.0	0.92	0.24
		2-spk	12.5	12.8	1.00	0.25
5	AV-DPRNN-ISAM*	2-spk	12.8	13.1	1.03	0.26
6	AV-DPRNN-ISAM†	1-spk	11.7	12.0	0.90	0.24
		2-spk	12.4	12.7	0.96	0.25
7	AV-TFGridNet [29]	1-spk	13.7	14.1	1.45	0.28
8	AV-TFGridNet-ISAM	1-spk	13.9	14.2	1.48	0.28
		2-spk	14.5	14.8	1.57	0.29

Table 2: *Results on VoxCeleb2 3-speaker mixture dataset. 1-spk means only the target speaker's face is visible, 2-spk includes the target and one random interfering speaker, and 3-spk includes all speakers.*

Sys.	Model	v-spks	SI-SNRi	SNRi	PESQi	STOIi
9	AV-DPRNN [28]	1-spk	10.6	11.1	0.37	0.29
10	AV-DPRNN-SE	3-spk	8.8	9.3	0.26	0.22
11	AV-DPRNN-ISAM	1-spk	11.5	12.0	0.45	0.30
		2-spk	11.9	12.3	0.46	0.31
		3-spk	13.3	13.7	0.61	0.34
12	AV-TFGridNet [29]	1-spk	14.2	14.6	0.89	0.36
13	AV-TFGridNet-ISAM	1-spk	14.3	14.7	0.93	0.37
		2-spk	14.8	15.2	0.96	0.38
		3-spk	15.6	15.9	1.08	0.39

Adam optimizer [39] and adopt a learning rate warmup strategy. Specifically, the learning rate (lr) is increased over the first $warmup_n = 15,000$ training steps according to the following formula:

$$\text{lr} = lr_{max} \cdot 0.001^{-1} \cdot 64^{-0.5} \cdot step_n \cdot warmup_n^{-1.5} \quad (3)$$

where $step_n$ is the current step number. The maximum learning rate lr_{max} is set to 0.001 for the AV-DPRNN series of models and 0.0001 for the AV-TF-GridNet series. After the warmup phase, the learning rate is halved if the validation loss does not improve for 6 consecutive epochs, and training stops if no improvement is observed within 10 subsequent epochs. All models are trained on four 80 GB A800 GPUs using the NVIDIA platform, with an effective batch size of 24 for the AV-DPRNN series and 16 for the AV-TF-GridNet series.

3.3. Network configuration

The hyperparameter settings for AV-DPRNN follow [27], while those for AV-TFGridNet follow [29]. For ISAM, the attention layer's embedding size matches the hidden dimension of the respective backbone, with a feedforward dimension twice the embedding size. The number of attention heads is set to 1. AV-DPRNN has 15.3 million (M) parameters, increasing to 15.5M with ISAM, while AV-TFGridNet grows from 20.8M to 21.0M with ISAM [2].

---

[2] The model and training scripts are available at `https://github.com/modelscope/ClearerVoice-Studio`

Table 3: *Results on MISP 2-speaker mixture dataset.*

Sys.	Model	v-spks	SI-SNRi	SNRi	PESQi	STOIi
14	AV-DPRNN [28]	1-spk	8.7	9.9	0.69	0.12
15	AV-DPRNN-ISAM	1-spk	8.7	9.9	0.70	0.11
		2-spk	11.4	12.4	0.86	0.16
16	AV-TFGridNet [29]	1-spk	10.4	11.6	1.03	0.15
17	AV-TFGridNet-ISAM	1-spk	10.5	12.0	1.08	0.15
		2-spk	13.3	14.5	1.26	0.19

Table 4: *Results on MISP 3-speaker mixture dataset.*

Sys.	Model	v-spks	SI-SNRi	SNRi	PESQi	STOIi
18	AV-DPRNN [28]	1-spk	6.7	7.8	0.26	0.12
19	AV-DPRNN-ISAM	1-spk	7.8	9.1	0.36	0.16
		2-spk	8.9	10.0	0.42	0.18
		3-spk	11.0	12.2	0.51	0.22
20	AV-TFGridNet [29]	1-spk	9.3	10.6	0.58	0.19
21	AV-TFGridNet-ISAM	1-spk	9.3	10.6	0.59	0.20
		2-spk	9.7	11.1	0.63	0.21
		3-spk	12.5	13.9	0.77	0.26

4. Result

We evaluate the extracted speech signals using two sets of metrics: SI-SNRi and SNRi (in dB) for signal quality, and PESQi and STOIi for perceptual quality and intelligibility. All metrics are computed relative to the unprocessed multi-talker speech signal, the higher the better. We primarily compare SI-SNRi, as other metrics show similar trends. Each trained system is uniquely identified by a system number (Sys.) in all tables.

4.1. Highly overlapped speech mixture

We first evaluate ISAM using the highly overlapped VoxCeleb2 mixture dataset. Table 1 presents results on 2-speaker mixtures. The baseline System 1, a vanilla AV-DPRNN, achieves an SI-SNRi of 11.5 dB. In contrast, System 2, which receives all faces as input and performs speech separation with two output streams, sees a drop in SI-SNRi to 10.7 dB, likely due to the network's limited capacity to process excessive information. System 3, which also takes all faces as input but extracts only the target speaker, performs better at 11.2 dB SI-SNRi but still lags behind system 1.

With our proposed ISAM dropout training, System 4 achieves 11.7 dB SI-SNRi when only the target speaker's face is available similarly to system 1, when an interfering face is also present (2-spk), SI-SNRi improves to 12.5 dB, demonstrating ISAM's effectiveness. As an upper bound, System 5 always employs ISAM and receives all speaker faces as input during training and inference, which achieves the highest 12.8 dB SI-SNRi. System 6 is a variation of System 4, where instead of ISAM dropout, the target branch always passes through ISAM even when no co-occurring face is available. Its performance is similar to System 4, but we opt for System 4 due to its lower parameter count and reduced computational cost thanks to the ISAM dropout.

System 7 is AV-TFGridNet baseline, achieving an SI-SNRi of 13.7 dB. System 8 is AV-TFGridNet with our ISAM. Due to the long training time of AV-TFGridNet models, all AV-TFGridNet-ISAM models in this paper are finetuned from the AV-TFGridNet baseline, whereas the AV-DPRNN-ISAM models are trained from scratch due to their simpler architecture and faster development cycle. For System 8, when only the target

Table 5: *Results on LRS2 2-speaker mixture dataset.*

Sys.	Model	v-spks	SI-SNRi	SNRi	PESQi	STOIi
1	AV-DPRNN [28]	1-spk	11.7	12.0	0.81	0.24
4	AV-DPRNN-ISAM	1-spk	11.9	12.2	0.85	0.24
		2-spk	12.8	13.2	0.93	0.25
7	AV-TFGridNet [29]	1-spk	14.4	14.7	1.44	0.28
8	AV-TFGridNet-ISAM	1-spk	14.4	14.7	1.45	0.28
		2-spk	15.0	15.4	1.55	0.29

Table 6: *Results on LRS2 3-speaker mixture dataset.*

Sys.	Model	v-spks	SI-SNRi	SNRi	PESQi	STOIi
9	AV-DPRNN [28]	1-spk	9.8	10.3	0.30	0.26
11	AV-DPRNN-ISAM	1-spk	10.9	11.5	0.38	0.29
		2-spk	11.2	11.6	0.37	0.30
		3-spk	13.2	13.6	0.52	0.33
12	AV-TFGridNet [29]	1-spk	14.0	14.4	0.79	0.36
13	AV-TFGridNet-ISAM	1-spk	13.9	14.4	0.81	0.36
		2-spk	14.4	14.8	0.85	0.37
		3-spk	15.7	16.1	1.00	0.39

Table 7: *Results on LRS3 2-speaker mixture dataset.*

Sys.	Model	v-spks	SI-SNRi	SNRi	PESQi	STOIi
1	AV-DPRNN [28]	1-spk	13.1	13.4	1.00	0.24
4	AV-DPRNN-ISAM	1-spk	13.3	13.5	1.04	0.24
		2-spk	14.2	14.5	1.15	0.25
7	AV-TFGridNet [29]	1-spk	16.2	16.5	1.73	0.27
8	AV-TFGridNet-ISAM	1-spk	16.2	16.5	1.77	0.27
		2-spk	16.9	17.1	1.86	0.28

Table 8: *Results on LRS3 3-speaker mixture dataset.*

Sys.	Model	v-spks	SI-SNRi	SNRi	PESQi	STOIi
9	AV-DPRNN [28]	1-spk	11.0	11.4	0.40	0.27
11	AV-DPRNN-ISAM	1-spk	12.1	12.5	0.50	0.30
		2-spk	12.5	12.8	0.51	0.30
		3-spk	14.7	15.0	0.70	0.34
12	AV-TFGridNet [29]	1-spk	15.2	15.6	0.99	0.36
13	AV-TFGridNet-ISAM	1-spk	15.1	15.5	1.02	0.35
		2-spk	15.9	16.1	1.08	0.37
		3-spk	17.2	17.4	1.26	0.38

speaker is visible, the SI-SNRi is 13.9 dB, but improves to 14.5 dB when all speakers are observed.

Table 2 presents results on VoxCeleb2 3-speaker mixtures. The AV-DPRNN baseline (System 9) achieves an SI-SNRi of 10.6 dB, while System 11, which incorporates ISAM, outperforms it in the 1-spk scenario with 11.5 dB SI-SNRi. This improvement is likely due to the enhanced performance in the 2-spk and 3-spk scenarios, which also benefits the 1-spk case. When one interfering speaker is observed (2-spk), SI-SNRi improves to 11.9 dB, and with all interference speakers present (3-spk), performance further increases to 13.3 dB. For AV-TFGridNet, the baseline (System 12) achieves 14.2 dB SI-SNRi. With ISAM, System 13 shows a steady improvement, increasing from 14.3 dB (1-spk) to 14.8 dB (2-spk) and reaching 15.6 dB (3-spk) as more interfering speakers are observed.

4.2. Sparsely overlapped speech mixture

Table 3 and Table 4 present results for 2-speaker and 3-speaker MISP mixtures, respectively. The results show a similar trend to those on the VoxCeleb2 mixture datasets, where ISAM consistently improves the performance of both AV-DPRNN and AV-TFGridNet across all conditions whenever any number of co-occurring faces are available.

We also observe that SI-SNRi on MISP is generally lower than on VoxCeleb2. This is because the sparsely overlapped mixtures pose a greater challenge: the network must jointly perform target speaker voice activity detection, and speaker separation when the target speaker is active. In contrast, in highly overlapped scenarios, the network only needs to focus on speaker separation, making the task comparatively easier. Nevertheless, ISAM provides even greater benefits in the sparsely overlapped scenario. For instance, on the 2-speaker mixtures, AV-TFGridNet-ISAM (System 8) shows a 0.6 dB SI-SNRi improvement from 1-spk to 2-spk on the VoxCeleb2 mixture dataset in Table 1, while AV-TFGridNet-ISAM (System 17) achieves a 2.8 dB SI-SNRi improvement on the MISP mixture dataset in Table 3. Similarly, on the 3-speaker mixtures, AV-TFGridNet-ISAM (System 13) improves SI-SNRi by 1.3 dB from 1-spk to 3-spk on VoxCeleb2 mixture dataset in Table 2, whereas AV-TFGridNet-ISAM (System 21) achieves

a 3.2 dB SI-SNRi improvement on MISP mixture dataset in Table 4. These results highlight the effectiveness of ISAM in handling sparsely overlapped mixtures, where leveraging co-occurring faces provides even greater advantages.

4.3. Cross-dataset evaluation

We also perform cross-dataset evaluations to assess the generalizability of ISAM. Table 5 and Table 6 present results for models trained on the VoxCeleb2 mixture but tested on the LRS2 mixture dataset. LRS2 is similar to VoxCeleb2, except that it contains only English speech, while VoxCeleb2 is multilingual. The results show that performance on LRS2 closely aligns with that on VoxCeleb2, with ISAM achieving similar performance gains over the baseline.

Additionally, Table 7 and Table 8 present results for models trained on VoxCeleb2 but tested on LRS3. Since LRS3 is a cleaner dataset than VoxCeleb2, the overall performance is higher. The performance gains of ISAM follow a similar trend to those observed on VoxCeleb2, further demonstrating its effectiveness across different datasets.

Across all datasets: VoxCeleb2, MISP, LRS2, and LRS3, we consistently observe that for 3-speaker mixtures, the performance gain from 1-spk to 2-spk is smaller than from 2-spk to 3-spk. This suggests that observing all faces provides the maximum performance improvement. Nevertheless, the 1-spk to 2-spk improvement remains significant, demonstrating that our model is robust to varying numbers of observed faces and effectively leverages available visual information.

5. Conclusion

In conclusion, we introduce the Inter-Speaker Attention Module (ISAM), a novel approach that enables audio-visual speaker extraction to utilize complementary speech activity cues from a flexible number of co-occurring on-screen faces. The ISAM not only enhances performance on highly overlapped speech mixtures but also achieves significant improvements on more challenging sparsely overlapped scenarios. Furthermore, our method demonstrates strong robustness and generalizability in cross-dataset evaluations, underscoring its effectiveness in real-world applications.

6. References

[1] J. Wang, Z. Pan, M. Zhang, R. T. Tan, and H. Li, "Restoring speaking lips from occlusion for audio-visual speech recognition," in *Proc. AAAI*, vol. 38, 2024.

[2] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[3] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[4] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[5] 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 *Proc. ICASSP*, 2023.

[6] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[7] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2840–2849, 2021.

[8] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. ICASSP*, 2017, pp. 246–250.

[9] T. von Neumann, K. Kinoshita, C. Boeddeker, M. Delcroix, and R. Haeb-Umbach, "SA-SDR: A novel loss function for separation of meeting style data," in *Proc. ICASSP*, 2022, pp. 6022–6026.

[10] Z. Pan, G. Wichern, F. G. Germain, K. Saijo, and J. Le Roux, "PARIS: Pseudo-AutoRegressIve siamese training for online speech separation," in *Proc. Interspeech*, 2024.

[11] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Learning speaker representation for neural network based multichannel speaker extraction," in *Proc. ASRU*, 2017, pp. 8–15.

[12] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. Interspeech*, 2019, pp. 2728–2732.

[13] 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*, vol. 13, no. 4, pp. 800–814, 2019.

[14] H. Sato, T. Ochiai, K. Kinoshita, M. Delcroix, T. Nakatani, and S. Araki, "Multimodal attention fusion for target speaker extraction," in *Proc. SLT*, 2021, pp. 778–784.

[15] C. Xu, W. Rao, E. S. Chng, and H. Li, "SpEx: Multi-scale time domain speaker extraction network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1370–1384, 2020.

[16] S. He, H. Li, and X. Zhang, "Speakerfilter: Deep learning-based target speaker extraction using anchor speech," in *Proc. ICASSP*, 2020, pp. 376–380.

[17] Z. Pan, R. Tao, C. Xu, and H. Li, "MuSE: Multi-modal target speaker extraction with visual cues," in *Proc. ICASSP*, 2021, pp. 6678–6682.

[18] J. Wu, Y. Xu, S. Zhang, L. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," in *Proc. ASRU*, 2019, pp. 667–673.

[19] Z. Pan, X. Qian, and H. Li, "Speaker extraction with co-speech gestures cue," *IEEE Signal Process. Lett.*, vol. 29, pp. 1467–1471, 2022.

[20] Z. Pan, W. Wang, M. Borsdorf, and H. Li, "ImagineNet: Target speaker extraction with intermittent visual cue through embedding inpainting," in *Proc. ICASSP*, 2023.

[21] E. Ceolini, J. Hjortkjær, D. D. Wong, J. O'Sullivan, V. S. Raghavan, J. Herrero, A. D. Mehta, S.-C. Liu, and N. Mesgarani, "Brain-informed speech separation for enhancement of target speaker in multitalker speech perception," *NeuroImage*, vol. 223, p. 117282, 2020.

[22] Z. Pan, M. Borsdorf, S. Cai, T. Schultz, and H. Li, "NeuroHeed: Neuro-steered speaker extraction using EEG signals," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2024.

[23] Z. Pan, G. Wichern, F. G. Germain, S. Khurana, and J. Le Roux, "NeuroHeed+: Improving neuro-steered speaker extraction with joint auditory attention detection," in *Proc. ICASSP*, 2024, pp. 11456–11460.

[24] Q. Wang, I. L. Moreno, M. Saglam, K. Wilson, A. Chiao, R. Liu, Y. He, W. Li, J. Pelecanos, M. Nika *et al.*, "VoiceFilter-Lite: Streaming targeted voice separation for on-device speech recognition," *Proc. Interspeech*, pp. 2677–2681, 2020.

[25] Z. Zhao, D. Yang, R. Gu, H. Zhang, and Y. Zou, "Target confusion in end-to-end speaker extraction: Analysis and approaches," *Proc. Interspeech*, 2022.

[26] K. Liu, Z. Du, X. Wan, and H. Zhou, "X-SepFormer: End-to-end speaker extraction network with explicit optimization on speaker confusion," in *Proc. ICASSP*, 2023, pp. 1–5.

[27] Z. Pan, M. Ge, and H. Li, "A hybrid continuity loss to reduce over-suppression for time-domain target speaker extraction," in *Proc. Interspeech*, 2022, pp. 1786–1790.

[28] ——, "USEV: Universal speaker extraction with visual cue," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 3032–3045, 2022.

[29] Z. Pan, G. Wichern, Y. Masuyama, F. G. Germain, S. Khurana, C. Hori, and J. Le Roux, "Scenario-aware audio-visual TF-Gridnet for target speech extraction," in *Proc. ASRU*, 2023.

[30] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. ICASSP*, 2020, pp. 46–50.

[31] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. ICASSP*, 2021, pp. 21–25.

[32] 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 *Proc. ICASSP*, 2023, pp. 1–5.

[33] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *Proc. ICASSP*, 2019, pp. 626–630.

[34] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," *Proc. Interspeech*, pp. 1086–1090, 2018.

[35] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.,*, 2018.

[36] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[37] H. Chen, H. Zhou, J. Du, C.-H. Lee, J. Chen, S. Watanabe, S. M. Siniscalchi, O. Scharenborg, D.-Y. Liu, B.-C. Yin, J. Pan, J.-Q. Gao, and C. Liu, "The first multimodal information based speech processing (MISP) challenge: Data, tasks, baselines and results," in *Proc. ICASSP*, 2022.

[38] Z. Pan, R. Tao, C. Xu, and H. Li, "Selective listening by synchronizing speech with lips," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1650–1664, 2022.

[39] D. P. Kingma and J. Ba, "Adam, a method for stochastic optimization," in *Proc. ICLR*, vol. 1412, 2015.