INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Rethinking the visual cues in audio-visual speaker extraction

Junjie Li[1], *Meng Ge*[2,4,*], *Zexu Pan*[3], *Rui Cao*[1], *Longbiao Wang*[1,*], *Jianwu Dang*[1], *Shiliang Zhang*

[1] Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[3] Institute of Data Science, National University of Singapore, Singapore
[4] Shenzhen Research Institute of Big Data, Shenzhen, China

mrjunjieli@tju.edu.cn,gemeng@nus.edu.sg,longbiao_wang@tju.edu.cn

Abstract

The Audio-Visual Speaker Extraction (AVSE) algorithm employs parallel video recording to leverage two visual cues, namely speaker identity and synchronization, to enhance performance compared to audio-only algorithms. However, the visual front-end in AVSE is often derived from a pre-trained model or end-to-end trained, making it unclear which visual cue contributes more to the speaker extraction performance. This raises the question of how to better utilize visual cues. To address this issue, we propose two training strategies that decouple the learning of the two visual cues. Our experimental results demonstrate that both visual cues are useful, with the synchronization cue having a higher impact. We introduce a more explainable model, the Decoupled Audio-Visual Speaker Extraction (DAVSE) model, which leverages both visual cues.

Index Terms: Visual cues, speaker extraction, identity, synchronization, decouple

1. Introduction

Speech is not only the most natural way of communication between humans, but also plays an indispensable role in human-computer interaction. Unfortunately, the speech of interest is always interfered by background noise and other speakers in the real world. While humans have the intrinsic ability to attend to the target speaker while ignoring other interference, also known as the cocktail party problem [1], machines have not been constructed to reach human standards.

The goal of speaker extraction is to separate target speech by filtering out environmental noise signals and other speakers' speech signals. It plays a critical role in speech preprocessing to facilitate downstream tasks, such as active speaker detection [2], speaker localization [3], speaker emotion analysis [4], and automatic speech recognition [5, 6]. In recent years, tremendous efforts have been made to improve the quality of separated speech, including techniques such as permutation invariant training [7], Conv-TasNet[8], dual-path RNN [9], SpEx+ [10], SpEx++ [11].

Human speech perception is essentially a multi-modal process. People not only listen to speech but also observe facial expressions and lip movements. According to neuroscience studies [12], visual inputs enhance people's ability to focus on the speaker of interest and reduce perceptual ambiguity in noisy environments. To mimic human perceptual processes, visual cues have been widely leveraged in recent studies [13, 14, 15, 16], which utilize visual cues as auxiliary information to extract corresponding target speech. Previous studies have reported great performance compared to audio-only speech separation [17, 18, 19, 20, 21], especially in noisy environments [22], attributed to the robustness of visual cues against acoustic noise.

There are two types of visual cues that are useful for speaker extraction: the speaker identity cue and the synchronization cue. The speaker identity cue can be learned from a single image [23, 24] or a video recording based on the studies of face-voice correlation. The synchronization cue is learned from a parallel video recording which contains speech-lip synchronization [25] of viseme-phoneme correlation [26, 27] information. Aldeneh et al. [28] have demonstrated that the performance varies depending on the articulation, indicating that the synchronization cue provides performance improvement. Another work [29] argues that auxiliary information is only beneficial for selecting the speaker of interest, indicating the effect of the speaker identity cue. Wang et al. [30] improve the performance by introducing auxiliary loss functions to model phonetic correlation between lip motion and phoneme, and speaker-identity correlation between timbre and facial attributes.

The state-of-the-art AVSE models usually employ a visual front-end to learn the visual cues. The visual front-end is either taken from part of a pre-trained network to extract low-level visual features or is trained end-to-end to optimize speaker extraction. Such training implicitly makes use of both the speaker identity and synchronization features. However, it is unclear which, or how much information is learned from each visual cue. Therefore, how to utilize visual cues remains an open question. We believe there is still room for improvement if we could explicitly decouple the learning of the two visual cues in one speaker extraction model which is the focus of our paper.

Different from [30], we propose two different training strategies to decouple the learning of two visual cues, namely the same-speaker aligned-visual training that is specialized in learning synchronization cue, and the different-speaker shuffled-visual training that is specialized in learning the speaker identity cue. Experimental results verify that both visual cues are useful, while the synchronization cue is clearly better. We also propose a Decoupled Audio-Visual Speaker Extraction model (DAVSE) to take advantage of both decoupled visual cues in one speaker extraction model. Our DAVSE outperforms baselines in terms of signal quality and perceptual evaluations. Our work provides a new sight into understanding the role of visual cues and presents views on how to improve the performance of AVSE.

2. Decoupled Audio-Visual Speaker Extraction Model

There are two types of visual cues that can be useful for speaker extraction: speaker identity cues and synchronization cues. These cues can help identify specific speakers and extract their

* Corresponding author.

10.21437/Interspeech.2023-2545

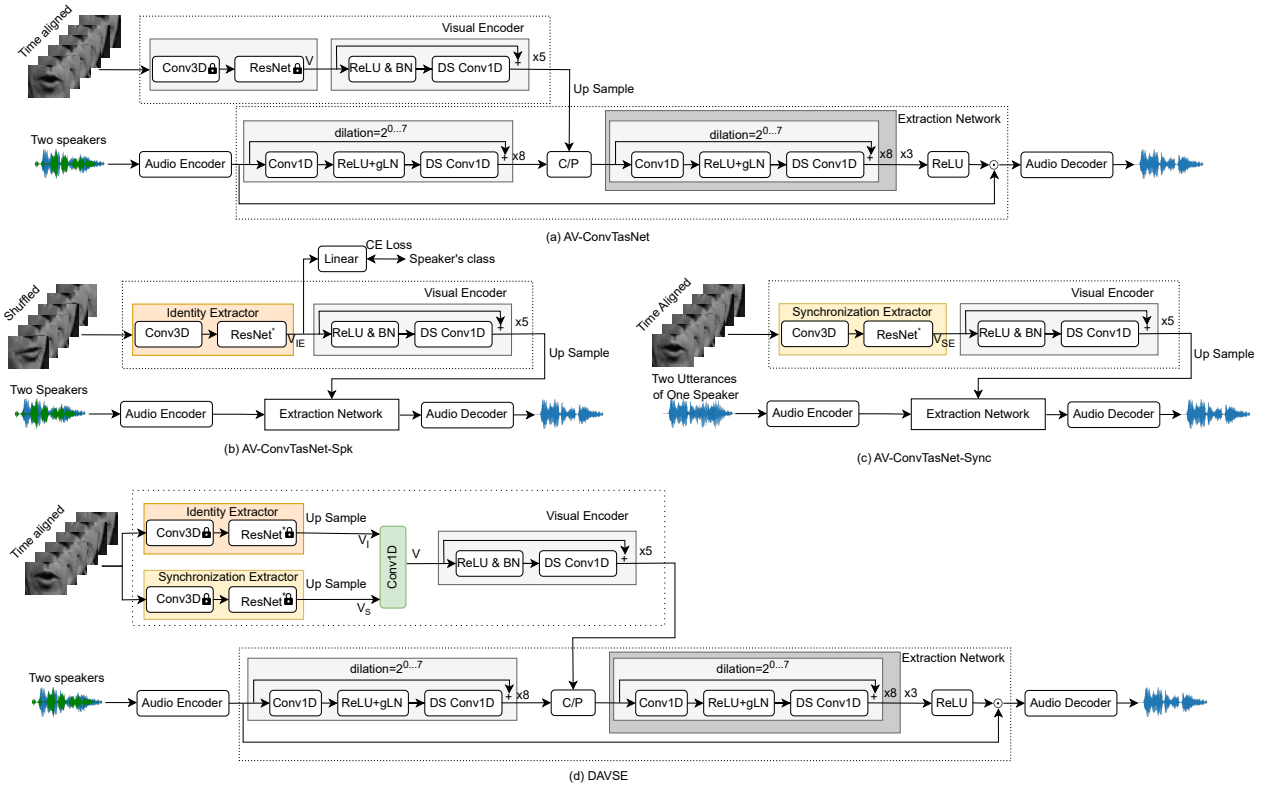

Figure 1: *Audio-visual speaker extraction models. (a) AV-ConvTasNet: Raw visual streams are used to extract target speech; hence, the speaker identity cue and synchronization cue are utilized implicitly. (b) AV-ConvTasNet-Spk: Only the speaker identity cue is utilized. (c) AV-ConvTasNet-Sync: Only the synchronization cue is utilized. (d) DAVSE: Both the speaker identity cue and synchronization cue are utilized explicitly.* $\odot$ *denotes point-wise multiplication, 'C/P' concatenates two input embeddings over the channel dimension and projects it to a lower dimension feature using Conv1D. The module with a lock symbol denotes that its weight is fixed during training.*

voices from mixed audio. Previous works typically concatenate audio and visual modalities in one model and train it end-to-end, thereby implicitly utilizing both cues. In this section, we propose two training strategies to decouple the speaker identity cue and synchronization cue from raw visual streams. Additionally, we design a Decoupled Audio-Visual Speaker Extraction model (DAVSE) that explicitly exploits both visual cues to improve speaker extraction performance.

2.1. Typical audio-visual speaker extraction model

A typical time-domain audio-visual speaker extraction model is exemplified by the AV-ConvTasNet [26], which contains four parts: a visual encoder, an audio encoder, an extraction network, and an audio decoder, as depicted in Figure 1(a). This model serves as our AV baseline for comparison.

The visual encoder has a 3D convolution (Conv3D) and a ResNet block followed by a video temporal convolutional block consisting 5 residual connected rectified linear unit (ReLU), batch normalization (BN) and depth-wise separable convolutional layers (DS-Conv1D) [26]. The weights of Conv3D and ResNet are pre-trained according to lip reading task, similar to the work [17]. The dimension of the output of ResNet $\boldsymbol{V}$ is 512.

The detailed architecture of audio encoder, audio decoder and extraction network can be found in work [26].

2.2. Decoupled training for speaker identity cue

To exploit speaker identity cue solely, we propose a different-speaker shuffled-visual training strategy, and name the model trained with this strategy as AV-ConvTasNet-Spk. The structure of AV-ConvTasNet-Spk is the same as AV-ConvTasNet except for the visual encoder. The identity extractor has a Conv3D and a ResNet* block. The dimension of output of ResNet* $\boldsymbol{V_{IE}}$ is only 256 here, as depicted in Fig. 1 (b).

To extract speaker identity feature, the cross-entropy (CE) loss is added for speaker classification. According to our experience, if only using separation loss, scale-invariant signal-to-noise ratio (SI-SNR) loss here, the model can not find a way to optimize. The training progress can be divided into two steps:

Step 1: The modules, Conv3D, ResNet* and Linear, are trained using CE loss. It is defined as :

$$\mathcal{L}_{CE} = -\sum_{l=0}^{L-1}\sum_{c=0}^{C-1} y_c log(softmax(\boldsymbol{W}V_{IE_l})) \qquad (1)$$

where C is the number of speakers in the training dataset. y_c is target speaker's class label. W is a learnable weight matrix for speaker classification. $V_{IE} \in R^{L\times N}$ is output feature of identity extractor. L and N are time and channel dimension, respectively. Therefore, the identity extractor can distinguish different speakers.

Step 2: We use the pre-trained identity extractor and fix these weights, and train other modules using speaker extraction

loss $\mathcal{L}_{SI-SNR}$. During training, the model takes speech mixed from two speakers and shuffled visual streams of target speaker. Because of the shuffled visual streams and pre-trained identity extractor, it forbids model to learn any synchronization cue, thus solely learning the speaker identity cue to distinguish different speakers.

2.3. Decoupled training for synchronization cue

To exploit synchronization cue solely, we propose a same-speaker aligned-visual training strategy, and name the model trained with this strategy as AV-ConvTasNet-Sync. It shares the same structure as AV-ConvTasNet-Spk except that it doesn't have speaker classification part. The dimension of output of ResNet* $\boldsymbol{V_{SE}}$ is only 256 here, as depicted in Fig. 1 (c).

During training, AV-ConvTasNet-Sync accepts speech mixed from different utterances of one speaker and time-aligned video and audio streams of the target speaker. The use of same-speaker speech mixture prevents the model from learning any identity cues, thereby allowing it to extract only the synchronization cue to extract the target speech.

2.4. DAVSE

To utilize both visual cues, we propose DAVSE [1]. Unlike AV-ConvTasNet having a single branch to model visual streams and exploit speaker identity and synchronization cues implicitly. We design two branches, identity extractor and synchronization extractor in a visual encoder, to extract speaker identity feature and synchronization feature. Identity extractor and synchronization extractor are fixed, and pre-trained from AV-ConvTasNet-Spk and AV-ConvTasNet-Sync, respectively, as depicted in Fig. 1 (d).

During training, DAVSE takes speech mixed from two speakers and time-aligned visual streams of target speaker. The outputs of identity extractor and synchronization extractor are concatenated along channel dimension, and then processed by 1D convolution to reduce dimension:

$$\boldsymbol{V_{IS}} = Concat(\boldsymbol{V_I}, \boldsymbol{V_S}) \tag{2}$$

$$\boldsymbol{V} = Conv1D(\boldsymbol{V_{IS}}, 1, 1) \tag{3}$$

The kernel size and stride are both set to 1. The channel dimension of $\boldsymbol{V_{IS}}$ and $\boldsymbol{V}$ are 512 and 256, respectively.

2.5. Loss function for speaker extraction

All models are trained using scale-invariant signal to noise ratio (SI-SNR) [31], which is defined as follows:

$$\begin{cases} s_{target} = \frac{\hat{s}^{\mathrm{T}} s}{\|s\|^2} s \\ e_{noise} = \hat{s} - s_{target} \\ SI\text{-}SNR(s, \hat{s}) = 10 log_{10} \frac{\|s_{target}\|^2}{\|e_{noise}\|^2} \\ \mathcal{L}_{SI\text{-}SNR}(s, \hat{s}) = -SI\text{-}SNR(s, \hat{s}) \end{cases} \tag{4}$$

where s and $\hat{s}$ denote the target speech and estimated speech, respectively

[1] https://github.com/mrjunjieli/DAVSE

3. Experiments

3.1. Lip Reading Sentences 3 (LRS3) dataset

LRS3 [32] is a large-scale audio-visual dataset that is obtained from TED and TEDx talks. There are 118,516 (408 h), 31,982 (30 h) and 1,321 (0.85 h) utterances in training, development (dev) and test sets, respectively. There are 5089 speakers in training set. The speakers in the train set and test set do not overlap.

3.2. Data preparation

The audio is sampled at 16k Hz, and corresponding video frames are sampled at 25 FPS. We use face recognition [2] algorithm to detect face for each frame and crop lip region from its face landmarks. Both face and lip images are used as visual input for speaker extraction task and resized to 112 * 112 pixels in greyscale.

To save computation resource, we pick 1,500 speakers and 1,000 speakers from training and dev sets, respectively. Among each speaker, short utterances (less than 4s) are dropped and long utterances are cut to 4$\sim$6s randomly. And test set is kept as the same as in LRS3. Finally, there are 41,560 utterances (1,500 speakers), 2,886 utterances (1,000 speakers) and 1,321 utterances (412 speakers) to simulate speech mixture in training, dev and test sets, respectively.

3.3. Data simulation

To decouple visual cues, we simulate three kinds of dataset [3]: different-speaker aligned-visual, different-speaker shuffled-visual and same-speaker aligned-visual. All speech mixtures are fully overlap.

different-speaker aligned-visual dataset: two audios from different speakers are mixed between -5 $\sim$ 10 dB in signal-to-interference ratio (SIR) [33]. And the visual reference is time aligned visual sequence of target speaker. Finally, we simulate 41,558 (about 50 h), 2,884 (about 5 h) and 1,320 utterances for training, dev and test set, respectively.

different-speaker shuffled-visual dataset: this is similar to the different-speaker aligned-visual dataset. Just the visual reference is shuffled each epoch during training.

same-speaker aligned-visual dataset: two different utterances from the same speaker are mixed between -5 $\sim$ 10 dB in signal-to-interference ratio (SIR). And the visual reference is time aligned visual sequence of target speaker. Finally, we simulate 41,338, 2,576 and 1,140 utterances for training, dev and test set, respectively.

3.4. Training details

We select Adam [34] as an optimizer. The initial rate is set to 10^{-3}. The learning rate is halved if the validation loss does not decrease for three epochs. The training process stops when validation loss does not decrease consecutively for six epochs or training epoch reaches 100.

4. Results

4.1. Comparison with baselines

Table 1 shows the performance of models under three simulated datasets. We evaluate the system's performance using SI-SNR

[2] https://pypi.org/project/face-recognition/

[3] https://github.com/mrjunjieli/LRS3_for_AVSS

Table 1: *SI-SNR (dB) and PESQ in a comparative study under different simulated datasets. 'D-S A-V' denotes different-speaker aligned-visual dataset. 'D-S S-V' denotes different-speaker shuffled-visual dataset. 'S-S A-V' denotes same-speaker aligned-visual dataset. 'Diff.' and 'Same' denote different and same gender mixtures, respectively.*

Methods	#Param		Visual Input	D-S A-V						D-S S-V		S-S A-V	
				SI-SNR			PESQ			SI-SNR	PESQ	SI-SNR	PESQ
	Total	Trainable		Diff.	Same	Avg.	Diff.	Same	Avg.	Avg.	Avg.	Avg.	Avg.
Mixture	\	\	\	-0.59	-0.06	-0.32	1.67	1.73	1.70	-0.32	1.70	0.24	1.79
AV-ConvTasNet [26]	16.99 M	5.8 M	lip	11.32	10.95	11.13	2.75	2.74	2.74	-5.82	1.52	8.97	2.57
AV-ConvTasNet-Sync	9.97 M	9.97 M	lip	9.79	10.45	10.13	2.59	2.69	2.64	-5.25	1.38	11.15	2.78
			face	10.55	10.93	10.74	2.68	2.75	2.72	-4.96	1.41	**11.82**	**2.86**
AV-ConvTasNet-Spk	9.97 M	9.97 M	lip	2.92	-0.80	1.03	1.88	1.62	1.75	0.86	1.73	-0.54	1.71
			face	5.29	-1.63	1.77	2.16	1.65	1.90	**1.51**	**1.88**	-2.28	1.57
DAVSE	15.32 M	4.88 M	lip	12.05	12.12	12.08	2.84	2.87	2.85	-4.13	1.65	10.73	2.72
			face	**12.77**	**12.70**	**12.73**	**2.93**	**2.95**	**2.94**	-4.41	1.63	10.58	2.76

and Perceptual Evaluation of Speech Quality (PESQ) [4] [35].

We evaluate the performance of AV-ConvTasNet-Sync and AV-ConvTasNet-Spk under 'D-S S-V' dataset and 'S-S A-V' dataset. AV-ConvTasNet-Sync and AV-ConvTasNet-Spk presents bad performance under 'D-S S-V' and 'S-S A-V', respectively, which indicates the effect of our proposed training strategies. AV-ConvTasNet-Sync only keeps synchronization cue and AV-ConvTasNet-Spk only keeps speaker identity cue.

Since synchronization cue is not affected by speaker information, AV-ConvTasNet-Sync shows similar performance under different-gender mixtures and same-gender mixtures. The performance of AV-ConvTasNet-Spk shows speaker identity cue is also useful to perform separation, especially under different-gender mixtures. These two models indicate that both two visual cues are useful for speaker extraction task, and synchronization cue is more important. The AV-ConvTasNet-Spk gets very bad performance when mixtures coming from the same gender, we guess that only using visual inputs are hard to distinguish speakers, and it biases the optimization of model towards easy mixture examples.

By utilizing synchronization cue and speaker identity cue explicitly, the proposed DAVSE presents performance improvement over other models. It proves the complementary effect of two visual cues compared to a single visual cue. And by modeling two visual cues explicitly, DAVSE also gets higher evaluation results compared to AV-ConvTasNet. The results of 'D-S S-V' and 'S-S A-V' also show that when visual streams are out of synchronization, it gets a very poor performance, indicating the importance of synchronization cue.

We also observe that face input contains more information in terms of not only speaker identity cue but also synchronization cue. Previous works [30, 26, 19] usually utilize lip streams to learn phonetic correlation between phoneme and lip motion. Our results indicate that facial expressions contain more information in term of phonetic correlation.

4.2. Visualization of visual embeddings

To visualize that DAVSE has learned a powerful visual embedding, Fig. 2 shows visual embeddings V of 9 random speakers from AV-ConvTasNet and DAVSE using uniform manifold approximation and projection (UMAP). Compared to embeddings of AV-ConvTasNet, the DAVSE's learned embeddings tend to distinguish not only speakers from different gender but also speakers from same gender. Therefore, DAVSE is easier to extract target speech from its interfering speech.

[4] https://github.com/vBaiCai/python-pesq

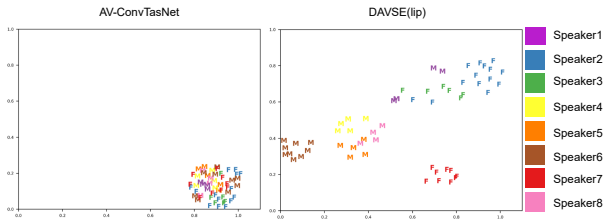

Figure 2: *The visual embeddings of 9 random speakers from test dataset visualized with UMAP [36]. **M** and **F** denote the male and female, respectively. To compare different embeddings on the same scale, we choose to normalize them using min-max normalization, which scales them to a range between 0.0 $\sim$ 1.0*

5. Conclusions

In this work, we explore the role of visual cues in audio-visual speaker extraction. We propose two different training strategies to decouple the learning of the synchronization and speaker identity cues. Experimental results show both visual cues are useful, while the synchronization cue is at the higher end. We also propose a more explainable model, named Decoupled Audio-Visual Speaker Extraction model (DAVSE), to take advantage of both decoupled visual cues in speaker extraction. Our DAVSE outperforms the baselines in terms of signal quality and perceptual evaluations.

6. Acknowledgements

This work is supported by 1) Huawei Noah's Ark Lab; 2) National Natural Science Foundation of China (Grant No. 62271432); 3) Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen (Grant No. B10120210117-KP02); 4) German Research Foundation (DFG) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen); 5) Alibaba Innovative Research Program.

7. References

[1] E. C. Cherry and W. Taylor, "Some further experiments upon the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 26, no. 4, pp. 554–559, 1954.

[2] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, "Is someone speaking? Exploring long-term temporal features for audio-visual active speaker detection," in *ACM MM*, 2021, pp. 3927–3935.

[3] X. Qian, M. Madhavi, Z. Pan, J. Wang, and H. Li, "Multi-target DoA estimation with an audio-visual fusion mechanism," in *ICASSP*, 2021, pp. 4280–4284.

[4] Z. Pan, Z. Luo, J. Yang, and H. Li, "Multi-modal attention for speech emotion recognition," in *INTERSPEECH*, 2020, pp. 364–368.

[5] A. Narayanan and D. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *TASLP*, vol. 22, no. 4, pp. 826–835, 2014.

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

[7] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *ICASSP*. IEEE, 2017, pp. 241–245.

[8] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

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

[10] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "SpEx+: A Complete Time Domain Speaker Extraction Network," in *Proc. Interspeech 2020*, 2020, pp. 1406–1410.

[11] ——, "Multi-stage speaker extraction with utterance and frame-level reference signals," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6109–6113.

[12] E. Z. Golumbic, G. B. Cogan, C. E. Schroeder, and D. Poeppel, "Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party"," *Journal of Neuroscience*, vol. 33, no. 4, pp. 1417–1426, 2013.

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

[14] J. Lee, S.-W. Chung, S. Kim, H.-G. Kang, and K. Sohn, "Looking into your speech: Learning cross-modal affinity for audio-visual speech separation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1336–1345.

[15] K. Li, F. Xie, H. Chen, K. Yuan, and X. Hu, "An audio-visual speech separation model inspired by cortico-thalamo-cortical circuits," *arXiv preprint arXiv:2212.10744*, 2022.

[16] Q. Liu, Y. Huang, Y. Hao, J. Xu, and B. Xu, "Limuse: Lightweight multi-modal speaker extraction," in *SLT*. IEEE, 2023, pp. 488–495.

[17] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *TPAMI*, 2018.

[18] J. Li, M. Ge, Z. Pan, L. Wang, and J. Dang, "VCSE: Time-domain visual-contextual speaker extraction network," in *INTERSPEECH*, 2022, pp. 906–910.

[19] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," in *CVPR*. IEEE, 2021, pp. 15 490–15 500.

[20] Z. Pan, R. Tao, C. Xu, and H. Li, "USEV: Universal speaker extraction with visual cue," *TASLP*, vol. 30, pp. 3032–3045, 2022.

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

[22] W. Wang, C. Xing, D. Wang, X. Chen, and F. Sun, "A robust audio-visual speech enhancement model," in *ICASSP*. IEEE, 2020, pp. 7529–7533.

[23] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang, "FaceFilter: Audio-visual speech separation using still images," in *INTERSPEECH*, 2020, pp. 3481–3485.

[24] L. Qu, C. Weber, and S. Wermter, "Multimodal target speech separation with voice and face references," *arXiv preprint arXiv:2005.08335*, 2020.

[25] Z. Pan, R. Tao, C. Xu, and H. Li, "Selective listening by synchronizing speech with lips," *TASLP*, vol. 30, pp. 1650–1664, 2022.

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

[27] 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.

[28] Z. Aldeneh, A. P. Kumar, B.-J. Theobald, E. Marchi, S. Kajarekar, D. Naik, and A. H. Abdelaziz, "On the role of visual cues in audiovisual speech enhancement," in *ICASSP*. IEEE, 2021, pp. 8423–8427.

[29] M. Elminshawi, W. Mack, S. Chakrabarty, and E. A. Habets, "New insights on target speaker extraction," *arXiv preprint arXiv:2202.00733*, 2022.

[30] X. Wang, X. Kong, X. Peng, and Y. Lu, "Multi-Modal Multi-Correlation Learning for Audio-Visual Speech Separation," in *Proc. Interspeech 2022*, 2022, pp. 886–890.

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

[32] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[33] 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.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, vol. 2. IEEE, 2001, pp. 749–752.

[36] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.