INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Multi-Stream Gated and Pyramidal Temporal Convolutional Neural Networks for Audio-Visual Speech Separation in Multi-Talker Environments

Yiyu Luo[1], *Jing Wang*[1], *Liang Xu*[1], *Lidong Yang*[2]

[1]School of Information and Electronics, Beijing Institute of Technology, China
[2]School of Information Engineering, Inner Mongolia University of Science and Technology, China
luoyiyu1104@gmail.com, wangjing@bit.edu.cn, xuliang981120@163.com, yld_nkd@imust.edu.cn

Abstract

Speech separation is the task of extracting target speech from noisy mixture. In applications like video telephones or video conferencing, lip movements of the target speaker are accessible, which can be leveraged for speech separation. This paper proposes a time-domain audio-visual speech separation model under multi-talker environments. The model receives audio-visual inputs including noisy mixture and speaker lip embedding, and reconstructs clean speech waveform for the target speaker. Once trained, the model can be flexibly applied to unknown number of total speakers. This paper introduces and investigates the multi-stream gating mechanism and pyramidal convolution in temporal convolutional neural networks for audio-visual speech separation task. Speaker- and noise-independent multi-talker separation experiments are conducted on GRID benchmark dataset. The experimental results demonstrate the proposed method achieves 3.9 dB and 1.0 dB SI-SNRi improvement when compared with audio-only and audio-visual baselines respectively, showing effectiveness of the proposed method.

Index Terms: audio-visual speech separation, cocktail party problem, temporal convolutional neural networks, gating mechanism, pyramidal convolution

1. Introduction

The goal of speech separation is to reconstruct clean speech while filtering out interference components, also known as the cocktail party problem [1]. Traditional speech separation algorithms can be classified into three categories, including signal processing-based like spectral subtraction [2] and Wiener filtering [3], decomposition-based like non-negative matrix factorization (NMF) [4] and rule-based like computational auditory scene analysis (CASA) [5]. Since speech separation problem is ill-posed, the performance of traditional algorithms is quite limited.

In recent years, with the development of machine learning, deep learning-based speech separation methods have made great progress, which usually treat speech separation as supervised problem, trying to learn nonlinear mapping function from noisy mixture to clean target speech representation. Due to the special pattern spectrogram demonstrates after time-frequency decomposition (STFT), most methods perform in frequency domain. Training targets include masking-based and mapping-based ones. Masking-based targets reflect the time-frequency energy ratios of clean speech to noisy interference, while mapping-based targets focus on directly recovering clean speech representation. Wang et al. [6] are the first to apply deep learning to speech separation, which models speech separation as a classification problem and uses DNNs to predict the ideal binary mask (IBM) [7]. Subsequently, various masks are proposed, including Ideal Ratio Mask (IRM) [8], complex Ideal Ratio Mask (cIRM) [9], etc. Multi-talker speech separation has long faced notorious label ambiguity and output dimension mismatch problem, which are caused by conflict between unknown number of unordered sound sources and ordered neural network outputs with certain dimensions. To address the above issues, Hershey et al. [10] train deep clustering embeddings (DPCL) for speech segmentation and separation. Yu et al. [11] propose a new training criterion named Permutation Invariant Training (PIT). Kolbæk et al. [12] enrich the original frame-level PIT to utterance-level PIT (uPIT).

For a long time, speech separation has been treated as audio-only issue. The McGurk effect [13] proves watching speaker's face movements can facilitate speech perception, illustrating deep interactions between audio and visual modality. Currently various visual cues have been introduced into speech separation task, including raw face or lip image pixels [14, 15, 16], face recognition embeddings [17], face landmarks [18], lip-reading embeddings [19, 20, 21], etc. Audio-visual separation methods usually introduce visual cues to facilitate audio-only separation methods. Gabbay et al. [14] propose an audio-visual convolutional autoencoder model, recovering enhanced spectrogram by lip regions. Afouras et al. [19] design a deep CNN-based speech enhancement network capable of extracting target speech given speaker lipreading embeddings. Ephrat et al. [17] apply face recognition embeddings and develop a deep speaker-independent audio-visual model on BLSTM network, achieving state-of-the-art separation performance even in challenging real-world scenarios. Morrone et al. [18] introduce face landmark motion features in LSTM-based model to generate time-frequency mask, realizing speaker-independent speech enhancement in multi-talker set-up though trained and evaluated on limited size datasets.

Recently, time-domain audio separation network Conv-Tasnet [22] has achieved great progress in end-to-end speech separation, surpassing previous frequency-domain methods. Wu et al. [21] then generalize the original audio-only Conv-Tasnet into audio-visual, gaining improvements over audio-only and other frequency-domain audio-visual separation models. This paper is partially built on audio-visual Conv-Tasnet [22, 21]. Conv-Tasnet mainly consists of stacked dilated temporal convolutional networks (TCN) [23], directly reconstructing clean speech waveform from noisy mixture. This paper focuses on improving the performance of TCN for audio-visual speech separation in multi-talker environments. The original TCN only contains single data stream, lacking capability of controlling information select and integrate compared with multi-stream gated networks. Besides, all filters in one specific temporal convolutional layer are of same kernel size, stride and dilation rate, resulting in same receptive fields thus failing to capture multi-scale features.

http://dx.doi.org/10.21437/Interspeech.2021-366

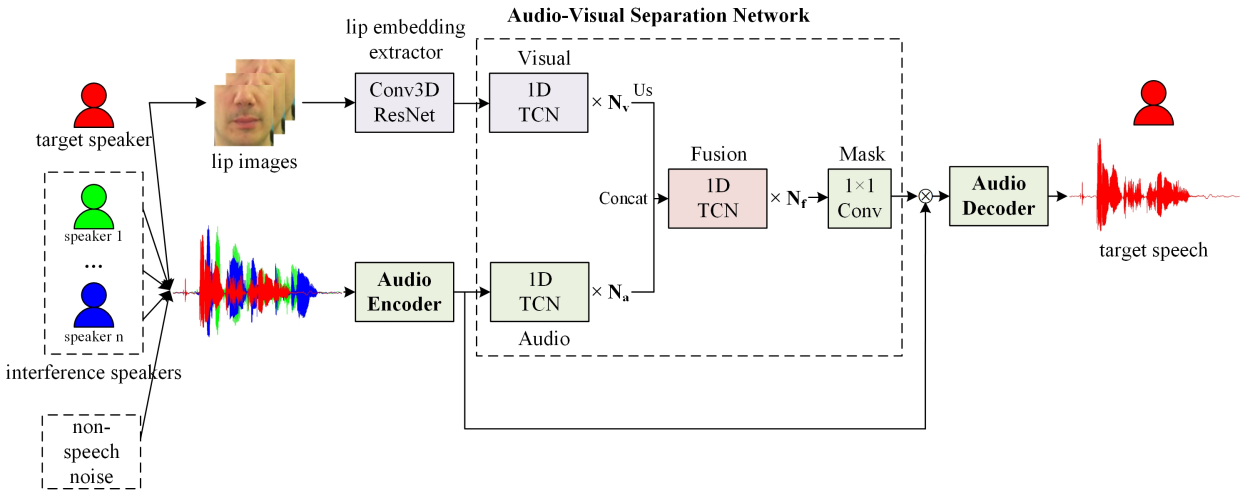

Figure 1: *System overview of the proposed audio-visual speech separation model. N_v, N_a and N_f denotes number of repeated 1D TCNs in visual, audio and fusion subnetworks, respectively. Us indicates upsampling visual features in the temporal dimension.*

The main contributions of our paper can be summarized as follows:

1. This paper introduces multi-stream gated TCN in audio-visual speech separation. By multiple individual data streams as well as information controlling gates, the method is capable of better modeling, selecting and integrating useful audio-visual representations for promoting speech separation performance.
2. The paper improves TCN with pyramidal convolutions. With pyramidal filters of different kernel size, the model is able to extract multi-scale audio-visual features within one convolutional layer, fully incorporating short-term and long-term information.

2. Architecture

2.1. System overview

As illustrated in Fig1, based on Conv-Tasnet, the proposed system mainly consists of three components: i) Audio encoder, which extracts audio representation from noisy mixture. ii) Audio-visual separation network, which generates a mask from the audio-visual representation for the target speaker. iii) Audio decoder, which reconstructs an enhanced speech waveform from the estimated mask and noisy mixture.

Our model receives a noisy mixture and lip regions from the target speaker, and predicts a corresponding clean speech waveform. Once trained, without knowing the total number of all speakers, the model can be flexibly applied to multi-talker separation by running multiple times with different audio-visual inputs of each individual speaker.

2.2. Visual representation

With regard to current audio-visual speech separation works, the most commonly used visual representations include raw face image pixels, face recognition embeddings and lip embeddings, etc. However, it's hard for model to learn audio-visual correlations directly from training on raw face images. Besides, face recognition embeddings focus more on face similarity instead of audio-visual inherent interactions. Lip embeddings are straightforwardly derived from lipreading task, thus capable of modeling correlations between lip movements and acoustic information, showing superiority in recent audio-visual speech separation task. In this paper, lip embeddings are adopted as visual representation.

The lip embeddings of the target speaker are extracted beforehand using a model pre-trained on lipreading task proposed in [24], with the implementation provided by [25]. The lipreading model consists of a 3D convolutional layer and a 18-layer ResNet [26], and generates a 512-dimensional lip embedding vector for every video frame. The input video segment is of 3-second uniform duration at 25 fps, thus leading to a 75×512 lip embedding vector containing speech-related visual features.

2.3. Audio encoder/decoder

Audio encoder extracts audio representation from noisy mixture waveform by performing 1D convolution operation, while audio decoder reconstructs enhanced speech waveform from the estimated mask and noisy mixture representation by performing 1D deconvolution operation. In this paper, 1D convolution is applied with number of filters $N = 512$, kernel size $K = 40$, stride $S = 20$.

2.4. Audio-visual separation network

Audio-visual separation network receives audio-visual representation including speaker lip embeddings and noisy audio feature, and predicts a mask for the corresponding target speaker. Visual, audio and fusion subnetworks are based on the identical fundamental building block, which are composed of $N_v = 1$, $N_a = 1$ and $N_f = 3$ repeated stacked 1D TCNs, respectively. Extra mask layer by 1×1 convolution is followed after fusion subnetwork.

The 1D TCN contains $X = 8$ blocks with different dilation rate 1, 2, .., 2^{X-1}, respectively, effectively enlarging receptive fields with reasonable model size. In this paper, as shown in Fig2, we investigate and compare three different structures of temporal convolutional block in audio-visual speech separation task, including basic, multi-stream gated and pyramidal temporal convolutional block. The basic block (Fig2.A) is built following Conv-Tasnet, consisting of two 1×1 convolutions and

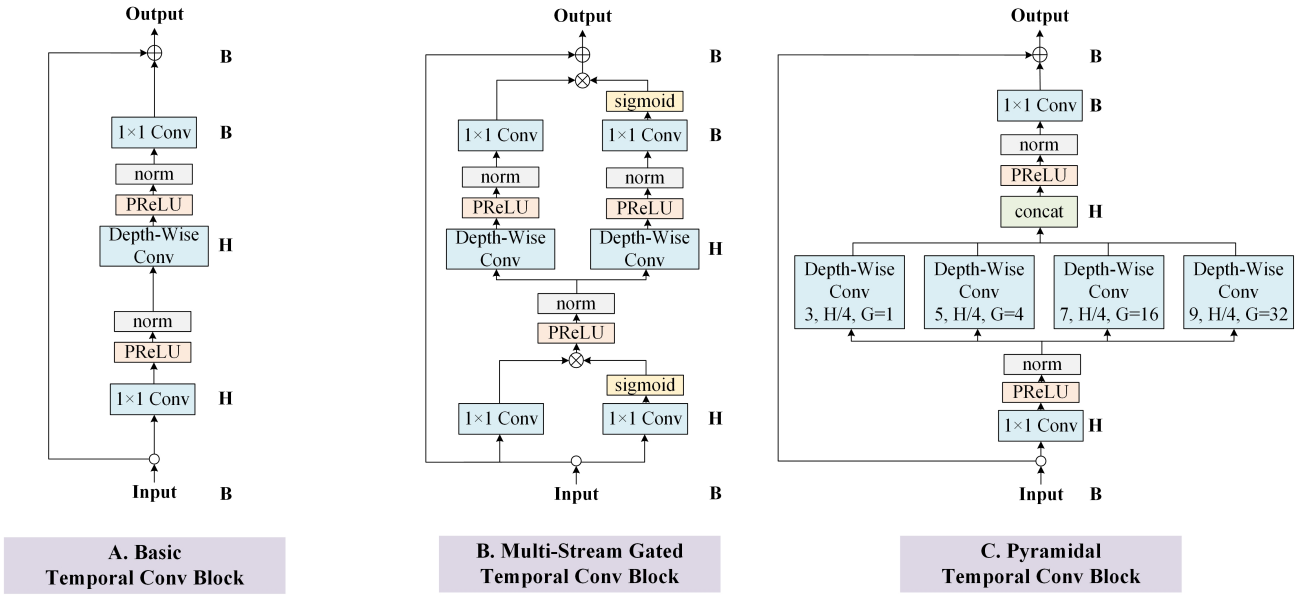

Figure 2: *Three different types of temporal convolutional block. B, H denote number of filters in corresponding 1D convolutional layers. G denotes convolutional groups. Both of the input and output vectors have B channels.* $\bigotimes$ *and* $\bigoplus$ *indicates point-wise multiplication and addition, respectively.*

one dilated depth-wise convolution as well as residual connection [26]. Parametric rectified linear unit (PReLU) [27] activation and global layer normalization (gLN) [28] are applied after convolutional layers. For all three blocks, we set number of convolutional filters $B = 128$, $H = 256$, default kernel size $P = 3$ except pyramidal block.

2.4.1. Multi-Stream Gated Block

Inspired by the success of gated TCN in audio-only separation [29], we introduce multi-stream gated block (Fig2.B) in audio-visual separation task. Compared with single-stream basic temporal block, the multi-stream gated block has stronger learning ability by two parallel data-processing streams. By sigmoid inflow and outflow gate functions, we assume multi-stream gated block is qualified for discarding irrelevant information while keeping enough necessary information for audio-visual speech separation.

2.4.2. Pyramidal Block

For basic temporal block, all filters in each convolutional layer share the same kernel size and stride, resulting in fixed receptive fields thus failing to extract multi-scale audio-visual features. Motivated by pyramidal convolution [30] in visual recognition, we design and introduce pyramidal temporal convolutional block (Fig2.C) in audio-visual speech separation task. Compared with basic block, the pyramidal block implements depth-wise convolution with filters of different kernel size and groups, the shape of which is similar to pyramidal. The pyramidal depth-wise convolution block contains 4 parallel depth-wise convolutions with different kernel size $P = 3, 5, 7, 9$, number of filters $H/4$ and groups $G = 1, 4, 16, 32$, respectively. Then a concatenation operation is performed at the channel dimension to obtain multi-scale audio-visual features, fully exploiting short-term and long-term dependencies in audio-visual speech separation.

3. Experimental Setup

3.1. Datasets

The model is trained and evaluated on GRID [31] audio-visual corpus dataset and MUSAN [32] noise dataset. For GRID dataset, 33 speakers (one missing due to technical oversight) are divided into disjoint sets of 25/4/4 for training/validation/test, respectively. Similarly, after discarding files shorter than 3-second, 775 non-speech noise clips in MUSAN dataset are split into disjoint sets of 575/100/100 for training/validation/test.

Synthetic data. Four types of synthetic samples are generated for multi-talker separation experiments, including 2-speaker mixture with/without background noise (2S Clean/ 2S Noisy) and 3-speaker mixture with/without background noise (3S Clean/ 3S Noisy). Each speech mixture is created by mixing target and interference speakers at a random signal-to-noise ratio (SNR) ranging from -5 dB to +5 dB. For training phase, noise is added to the speech mixture with a multiplicative factor of 0.2. As for test phase, noise is added to the speech mixture at a random SNR between -5 dB and +5 dB if necessary. The model is trained on 2S Noisy and tested on all four types of synthetic samples.

3.2. Data pre-processing

Video. The raw video from GRID dataset is of 3-second duration at 25 fps, containing a whole front face of the corresponding talking speaker. Dlib face landmark detector [33] implemented on [34] is used to locate and crop lip regions. Then, the mouth areas are resized to 88×88 size and converted into gray images for further lip embedding extraction.

Audio. All speech and noise samples are clipped to 3-second and resampled to 8000 Hz. To prevent loud voice from overwhelming soft one, Z-score normalization with 0 mean and standard deviation of 1 is performed to all speech and noise files.

Table 1: *Experimental results on speaker-independent and noise-independent multi-talker separation. SI-SNRi and PESQ are adopted as evaluation metrics.*

	2S clean		2S Noisy		3S clean		3S Noisy	
	SI-SNRi	**PESQ**	**SI-SNRi**	**PESQ**	**SI-SNRi**	**PESQ**	**SI-SNRi**	**PESQ**
AO Conv-Tasnet	9.02	2.66	9.20	2.07	-0.29	1.97	3.54	1.68
AV Conv-Tasnet	9.83	2.88	10.40	2.25	5.27	2.25	7.53	1.86
AV GTCN	10.99	3.00	11.59	2.35	5.72	2.28	8.16	1.89
AV PyTCN	11.01	3.02	11.92	2.37	6.28	2.35	8.73	1.92

3.3. Training

The training target of the model is scale-invariant source-to-noise ratio (SI-SNR) [35], which improves the original source-to-distortion ratio (SDR) [36] metric. SI-SNR is defined as:

$$\text{SI-SNR}(s, \hat{s}) = \frac{|\alpha s|^2}{|\alpha s - \hat{s}|^2} \text{ for } \alpha = \arg\min_{\alpha} |\alpha s - \hat{s}|^2 \quad (1)$$

s and $\hat{s}$ denote reference and estimated speech signal respectively, where optimal scaling factor $\alpha = \hat{s}^T s / \|s\|^2$.

The model is implemented in PyTorch and trained by maximizing SI-SNR metric between the reference and estimated speech waveforms. Adam optimizer [37] is used with an initial learning rate of 1e-3 which is reduced by half when learning process gets stuck.

3.4. Evaluation Metrics

In our experiments, SI-SNR improvement (SI-SNRi) and Perceptual Evaluation of Speech Quality (PESQ) [38] are adopted as evaluation metrics. SI-SNRi is defined as:

$$\text{SI-SNRi} = \text{SI-SNR}(s, \hat{s}) - \text{SI-SNR}(s, mix) \quad (2)$$

s, $\hat{s}$ and mix denote clean reference, estimated speech signal and noisy mixture, respectively. Higher SI-SNRi and PESQ values indicate better separation performance and speech quality.

3.5. Model configurations

To test the effectiveness of the proposed method, multi-talker speech separation experiments are conducted with the following models, including one audio-only and three audio-visual models. The experiments are both speaker-independent and noise-independent, which means the speaker identities and noise types during test phase are unseen in training and validation phase. Four test setups are considered, including 2-speaker/3-speaker mixture with/without background noise.

AO Conv-Tasnet. Audio-only Conv-Tasnet is adopted as audio-only baseline. Different from our audio-visual separation networks with single output, the output number of AO Conv-Tasnet is the same as the total number of speakers. PIT method [11] is used while training AO Conv-Tasnet model.

AV Conv-Tasnet. Audio-visual Conv-Tasnet with basic temporal convolutional blocks.

AV GTCN. Audio-visual Conv-Tasnet with multi-stream gated temporal convolutional blocks.

AV PyTCN. Audio-visual Conv-Tasnet with pyramidal temporal convolutional blocks. The only difference between the three audio-visual separation models is the structure of temporal convolutional blocks.

4. Results

Tab1 shows quantitative multi-talker separation results on SI-SNRi and PESQ metrics of the above four models.

For AV Conv-Tasnet model and AO Conv-Tasnet, audio-visual model performs better in terms of all indexes, indicating incorporating visual modality can benefit speech separation. Comparing separation results under 2S Clean and 3S Clean conditions, when total number of speakers increases, the SI-SNRi metric of audio-only model drop more significantly (-9.31 dB) than audio-visual model (-4.56 dB), showing visual cues are discriminant features for multi-talker separation.

Since SI-SNRi indicate the SI-SNR improvement after speech separation, for same total number of speakers, SI-SNRi under noisy conditions are more obvious than clean ones due to lower noisy SI-SNR before separation. On the contrary, PESQ in noisy conditions is lower than clean ones, because noise interference degrades speech quality.

With regard to our proposed AV GTCN and AV PyTCN models, the two models perform almost the same under 2-speaker mixture, but AV PyTCN model is slightly superior to AV GTCN model under 3-speaker mixture, showing the significance of multi-scale features extracted by pyramidal convolutions. The two our proposed models generally gain 3.93 dB SI-SNRi and 0.30 PESQ improvement than audio-only baseline, as well as 1.04 dB SI-SNRi and 0.09 PESQ than audio-visual baseline. In further experiments, AV GTCN and AV PyTCN models still show an advantage over audio-visual baseline with increased comparative parameters, proving the effectiveness of our proposed method for multi-talker speech separation task.

5. Conclusion

In this paper, we investigate and improve time-domain audio separation network Conv-Tasnet for audio-visual separation in multi-talker environments. The proposed model accepts multi-modal inputs including lip embeddings and noisy mixture, and predicts enhanced waveform for the target speaker. Two modification mechanisms are introduced including multi-stream gated and pyramidal temporal convolutional blocks. Experimental results on GRID benchmark dataset illustrate our proposed models achieve 13% relative improvement on SI-SNRi metric compared with audio-visual baseline model, showing superiority of the proposed method.

6. Acknowledgements

We would like to thank Pingchuan Ma for offering the pre-trained model of the lipreading network. This work was supported by National Natural Science Foundation of China (Grant No. 62071039 and 61620106002).

7. References

[1] 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.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[3] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1218–1234, 2006.

[4] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[5] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[6] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[7] G. Hu and D. Wang, "Speech segregation based on pitch tracking and amplitude modulation," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*. IEEE, 2001, pp. 79–82.

[8] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[9] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[10] 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.

[11] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[12] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[13] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, p. 746, 1976.

[14] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," *arXiv preprint arXiv:1711.08789*, 2017.

[15] M. Gogate, A. Adeel, R. Marxer, J. Barker, and A. Hussain, "Dnn driven speaker independent audio-visual mask estimation for speech separation," *arXiv preprint arXiv:1808.00060*, 2018.

[16] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 631–648.

[17] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *arXiv preprint arXiv:1804.03619*, 2018.

[18] G. Morrone, S. Bergamaschi, L. Pasa, L. Fadiga, V. Tikhanoff, and L. Badino, "Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6900–6904.

[19] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *arXiv preprint arXiv:1804.04121*, 2018.

[20] T. Afouras, J. S. Chung, and A. Zisserman, "My lips are concealed: Audio-visual speech enhancement through obstructions," *arXiv preprint arXiv:1907.04975*, 2019.

[21] J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," *arXiv preprint arXiv:1904.03760*, 2019.

[22] 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.

[23] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 47–54.

[24] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with lstms for lipreading," *arXiv preprint arXiv:1703.04105*, 2017.

[25] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6548–6552.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[28] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[29] Z. Shi, H. Lin, L. Liu, R. Liu, J. Han, and A. Shi, "Deep attention gated dilated temporal convolutional networks with intra-parallel convolutional modules for end-to-end monaural speech separation." in *Interspeech*, 2019, pp. 3183–3187.

[30] I. C. Duta, L. Liu, F. Zhu, and L. Shao, "Pyramidal convolution: rethinking convolutional neural networks for visual recognition," *arXiv preprint arXiv:2006.11538*, 2020.

[31] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[32] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[33] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.

[34] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1867–1874.

[35] 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.

[36] 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.

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

[38] 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 *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.