DEEP AUDIO-VISUAL SPEECH SEPARATION WITH ATTENTION MECHANISM

Chenda Li, Yanmin Qian

MoE Key Lab of Artificial Intelligence
SpeechLab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
lichenda1996@sjtu.edu.cn, yanminqian@sjtu.edu.cn

ABSTRACT

Previous work shows that audio-visual fusion is a practical approach to deal with the speech separation task in the cocktail party problem. In this paper, we explore a better strategy to utilize visual representations with the attention mechanism. Compared to the previous baseline only using one visual stream of the target speaker, both speaker-dependent visual streams in the mixed audio are fed into the model, and it also predicts two separated speech streams simultaneously. To further enhance the performance, the attention mechanism is designed on the audio-visual speech separation architecture. The results show that the proposed approach works well in audio-visual speech separation. Our best model achieves an obvious and consistent improvement in speech separation when compared to the traditional method only using the target speaker visual stream.

Index Terms— Audio-Visual, Speech Separation, Multimodal, Attention mechanism

1. INTRODUCTION

Speech separation and enhancement is one of the most important key technologies to solve the cocktail party problem [1, 2], where overlapped speech often exists in this environment. In such a scenario, separating the target speaker's speech from a noisy overlapped speech mixed with other speakers is an interesting and challenging problem. In recent years, various speech separation technologies have been proposed, such as deep clustering (DPCL) [3, 4, 5], permutation invariant training (PIT) [6, 7, 8] and beamforming based methods [9, 10]. Depending on the number of microphones, these methods can be grouped into multi-channel and single-channel based approaches. Usually the multi-channel approaches have better performance, since it can use additional information such as spatial location.

All of the above methods utilize only the audio signal for speech separation. Recently, multi-modality receives more and more interests, and other additional information rather than the audio is also explored, such as visual information. More and more audio-visual based speech processing system is developed for some speech-related tasks. Previous researches have successfully introduced visual information into the speech recognition task [11, 12]. For audio-visual speech enhancement and separation, there are also some preliminary works [13, 14]. The work in [14] shows that on signal-channel speech separation, the audio-visual speech separation models outperform the audio-only models. With the growth of audio-visual datasets and computational resources [15, 16, 17], it is more important to use deep neural networks to combine both audio and visual information together to address some hard speech-related problems, which can make it more robust when facing more complex scenarios.

In this paper, we explore a better strategy for audio and visual information fusion in speech separation task with the attention mechanism. Firstly, an audio-visual speech separation baseline model is built following the process described in [13] where only one visual stream of the target speaker is used in the architecture. Secondly, we extends this basic structure to utilize the visual streams of both speakers in the mixed speech for model optimization. The results show that visual information which belongs to the interference speaker is also useful for extracting the target speaker stream. Considering that the attention mechanism [18] has shown its effectiveness in various tasks, including natural language processing [19, 20] and speech recognition [21, 22], an attention mechanism is then designed to make a better utilization on both speaker visual representations in the mixed audio. To the best of our knowledge, few work has been done on using attention mechanisms for audio-visual speech separation. Our experiments show that the proposed architecture can get a consistent performance improvement on speech separation when compared to the traditional method.

The rest of this paper is organized as follows. Section 2 revisits the baseline model only using the target speaker's visual stream, and Section 3 describes our proposed new architecture using the visual information of both speakers in the mixed speech and further enhanced with the attention module. In section 4, experimental results are compared and analyzed. Finally, a conclusion is given in section 5.

2. BASELINE USING ONE TARGET SEPAKER VISUAL STREAM

The magnitude sub-network proposed in [13] is implemented with MXNet [23] as our baseline model, which only utilizes one visual stream of the target speaker in the architecture. As Fig.1 shows, the magnitude network takes the visual representation[1] $\mathbf{V}$ of the target speaker and the noisy audio representation $\mathbf{M}$ of the mixed speech as inputs.

The $ResNets$ [24] in Fig.1 comprises a stack of basic blocks. Each basic block consists of a 1D convolution layer with a resid-

Yanmin Qian is the corresponding author.

This work has been supported by the China NSFC project No. U1736202. Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University

[1] For convenience, we do not distinguish the notation of high-level representations in different stages and original input representations of the same stream.

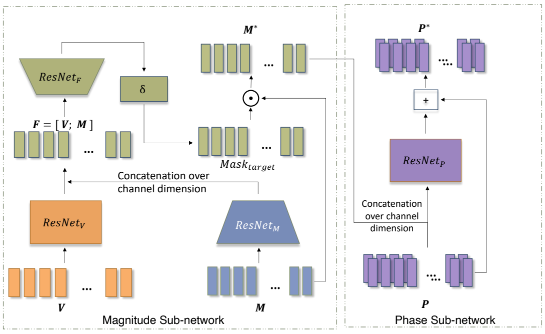

Fig. 1. The baseline model. The left part is magnitude sub-network; The right part is phase sub-network; $\mathbf{V}$ is visual representation; $\mathbf{M}$ is noisy audio representation; $\mathbf{F}$ is fusion representation; $Mask_{target}$ is predicted magnitude mask; $\mathbf{M}^*$ is predicted magnitude spectrogram; δ is sigmoid function; $\mathbf{P}$ is noisy phase spectrogram; $\mathbf{P}^*$ is predicted phase spectrogram.

ual connection, a ReLU activation layer, and a batch normalization layer. Some of the basic blocks in the $ResNet$ contain an extra up-sampling or down-sampling layer. Since the length of input audio representation $\mathbf{M}$ is four[2] times longer than the visual representation $\mathbf{V}$, two basic blocks in the $ResNet_M$ contain a down-sampling layer with factor 2. There is no down-sampling layer in $ResNet_V$. Therefore, after the processing of $ResNet_V$ and $ResNet_M$ respectively, the audio representation $\mathbf{M}$ and the visual representation $\mathbf{V}$ have the same length. Then, these two streams of representations are concatenated over the channel dimension to get the fusion representation $\mathbf{F} = [\mathbf{V}; \mathbf{M}]$.

The fusion representation $\mathbf{F}$ is passed to the fusion stack $ResNet_F$. Two basic blocks in the $ResNet_F$ contain a up-sampling layer with factor 2, so it up-samples the fusion representation $\mathbf{F}$ to the same length as the input audio representation. The last convolution layer in the $ResNet_F$ projects the dimension of fusion representation into the same as the noisy magnitude spectrogram. Then, the fusion representation is activated by a sigmoid function layer to obtain a magnitude mask for the target speaker. The values in the mask are between 0 and 1. The noisy magnitude spectrogram $\mathbf{M}$ is element-wise multiplied with the predicted magnitude mask to get the predicted magnitude spectrogram of the target speaker:

$$\mathbf{M}^* = \mathbf{M} \odot Mask_{target} \tag{1}$$

Previous research [25] shows that L1 loss is better than L2 loss when separating the target speaker, so L1 loss is chosen in the network training. Denote the reference of the target magnitude spectrogram as $\hat{\mathbf{M}}$, the network is optimized:

$$\mathcal{L}_{magnitude} = \|\mathbf{M}^* - \hat{\mathbf{M}}\|_1 \tag{2}$$

The phase sub-network takes the predicted magnitude spectrogram and noisy phase representation (sine and cosine value) as input. The two streams are concatenated together over the channel dimension and then processed by $ResNet_P$. The $ResNet_P$ outputs phase residual, and the phase residual is added to the noisy phase. After L_2-normalization, the predicted phase is obtained. Denote $\mathbf{P}$ as the noisy phase spectrogram, and $\mathbf{P}^*$ as the predicted phase spectrogram, more formally, the procedure can be presented as:

$$\mathbf{P}^* = \frac{(ResNet_P(\mathbf{P}) + \mathbf{P})}{\|(ResNet_P(\mathbf{P}) + \mathbf{P})\|_2} \tag{3}$$

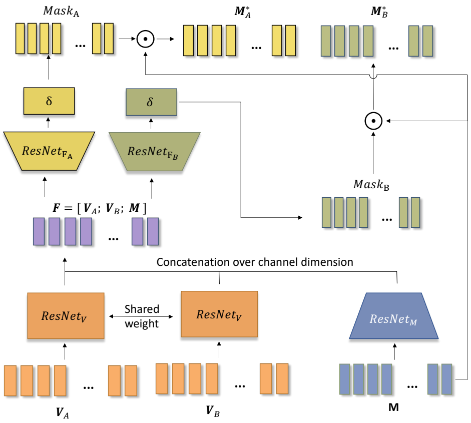

Fig. 2. The magnitude network of the proposed audio-visual separation model. $\mathbf{V}_A$ and $\mathbf{V}_B$ are visual representations; $\mathbf{M}$ is noisy audio representation; $\mathbf{F}$ is fusion representation; $Mask_A$ and $Mask_B$ are predicted magnitude masks; $\mathbf{M}^*_A$ and $\mathbf{M}^*_B$ are predicted magnitude spectrograms; δ is a sigmoid function.

Cosine similarity loss is used to optimize the phase sub-network. We do not make new modification on the phase sub-net work in this paper, and just borrow the same structure from the previous work [13].

3. PROPOSED DEEP AUDIO-VISUAL SPEECH SEPARATION WITH ATTENTION

3.1. Audio-Visual Speech Separation with both Visual Streams in the Mixed Speech

The baseline architecture only uses the visual stream of the target speaker for target speaker separation. In this work, we extends this basic structure to utilize the visual streams from both the target speaker and the interference speaker. Considering a mixed audio from speaker A and B, when we extract speech A from the mixed audio, the information from visual representation B is not being used in the baseline architecture. However the visual information from the interference speaker may be also useful, so it is explored here.

Under the assumption that at least two speakers' visual representations are available, we proposed the audio-visual speech separation model. As Fig. 2 shows, $ResNet_V$ and $ResNet_M$ are the same as those in the baseline model. The two streams of visual representation $\mathbf{V}_A$ and $\mathbf{V}_B$ are processed by the weight-shared $ResNet_V$. All three streams of representation are then concatenated together over the channel dimension. Then, the fusion representation $\mathbf{F} = [\mathbf{V}_A; \mathbf{V}_B; \mathbf{M}]$ is processed by two fusion nets $ResNet_{F_A}$ and $ResNet_{F_B}$ at the same time to predict both masks for both speaker in the mixed speech. These two fusion ResNets have the same structure as $ResNet_F$ in the baseline model, but each of them has its own parameters.

Denote the references of target magnitude spectrograms for both speakers as $\hat{\mathbf{M}}_A$, $\hat{\mathbf{M}}_B$, and the predicted magnitude spectrogram as

[2] see details in section 4.2

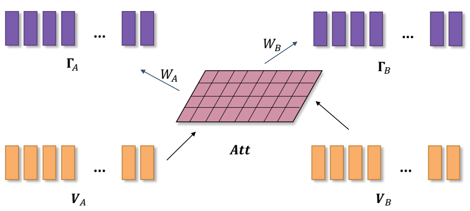

Fig. 3. The proposed attention mechanism. $\mathbf{V}_A$ and $\mathbf{V}_B$ are visual representations for different speakers; $\mathbf{\Gamma}_A$ and $\mathbf{\Gamma}_B$ are attended new feature maps; $\mathbf{Att}$ is attention weight matrix.

$\mathbf{M}^*_A$ and $\mathbf{M}^*_B$, the optimization objective can be:

$$\mathcal{L} = \frac{\|\mathbf{M}^*_A - \hat{\mathbf{M}}_A\|_1 + \|\mathbf{M}^*_B - \hat{\mathbf{M}}_B\|_1}{2} \tag{4}$$

3.2. Attention Mechanism for Audio-Visual Speech Separation

In section 3.1, we have introduced the proposed audio-visual speech separation model with both target and interfered speakers' visual streams. Our experiment shows that compared to the baseline, introducing the visual representation of the interference speaker can bring further improvement when separating the target speech, which will be shown in the next section. Furthermore, we have explored to incorporate the attention mechanism to the audio-visual speech separation model, in order to help the model focus more on the differences and similarities between the visual representations of different speakers.

The work in [20] shows that attending feature maps generated from two sequences works well in convolution neural network. Inspired by that, we propose our attention based audio-visual speech separation model. As Fig.3 shows, we denote visual representation A and B processed by previous visual ResNet in Fig.2 as $\mathbf{V}_A, \mathbf{V}_B \in \mathbf{R}^{c\times t}$. The length of $\mathbf{V}_A$ and $\mathbf{V}_B$ is t, and c is each frame's number of dimensions. The attention weight matrix $\mathbf{Att} \in \mathbf{R}^{t\times t}$ can be represented as:

$$\mathbf{Att}_{i,j} = attenion\_score(\mathbf{V}_A[:,i], \mathbf{V}_B[:,j]) \tag{5}$$

For scaled dot-product attention:

$$attenion\_score(\mathbf{V}_A[:,i], \mathbf{V}_B[:,j]) = \frac{\mathbf{V}_A[:,i]\cdot\mathbf{V}_B[:,j]}{\sqrt{c}} \tag{6}$$

The i-th row in the attention weight matrix $\mathbf{Att}$ denotes the attention distribution of i-th frame in visual representation $\mathbf{V}_A$ with respect to visual representation $\mathbf{V}_B$, and vice versa. Thus, the attention feature maps Γ_A and Γ_B are generated from matrix $\mathbf{Att}$ through learnable fully connected layer $\mathbf{W}_A$ and $\mathbf{W}_B$[3]:

$$\begin{aligned}\mathbf{\Gamma}_A &= \mathbf{W}_A \cdot \mathbf{Att}^T \\ \mathbf{\Gamma}_B &= \mathbf{W}_B \cdot \mathbf{Att}\end{aligned} \tag{7}$$

Finally, for stream A, the fusion representation $\mathbf{F}_A$ consists of two streams of visual representation $\mathbf{V}_A$, $\mathbf{V}_B$, the noisy audio representation $\mathbf{M}$ and the attention feature map $\mathbf{\Gamma}_A$. By concatenating those representations over the channel dimension, the fusion representation for stream A can be presented as $\mathbf{F}_A = [\mathbf{V}_A; \mathbf{V}_B; \mathbf{M}; \mathbf{\Gamma}_A]$. For stream B, $\mathbf{F}_B = [\mathbf{V}_A; \mathbf{V}_B; \mathbf{M}; \mathbf{\Gamma}_B]$. Similar to the model proposed in section 3.1, the fusion representation is then passed to the same pipeline, which are the same as the model described in the section 3.1.

4. EXPERIMENTS

4.1. Dataset

Models are trained on the LRS2 dataset [16], which consists of spoken sentences and corresponding videos from BBC television. The audio and video in the dataset are already synchronized. The list for splitting the dataset is provided, and it is divided into training, validation and testing set by broadcasting date, so there is no overlapping between the sets. There are about 142k utterances in the training set, about 1k in the validation and testing respectively. Videos in the dataset are all at 25fps, and audios are recorded at 16kHz sample rate.

In order to show the generalization of the proposed models, we also evaluate our model trained on LRS2 on one subset of VoxCeleb2 dataset [17]. We randomly selected 2000 samples in the VoxCeleb2 dataset for generalization testing.

4.2. Audio and Visual Representation

The data preparing procedure is similar to that in the previous work [13]. Each sample in the dataset is converted to visual and audio representations in advance.

Visual representation: A lip reading model is firstly trained on LRW dataset [15] following the recipe in [12, 11]. The visual-only model[4] in [11] is used for the lip reading task. The model we trained reaches 75.4% accuracy on the LRW validation set. Then the 18-layer 3D ResNet front end of the lip reading model is used to extract 512-dimensional features for each video frame. The visual features are clipped or padded to a fixed length of 60. In Fig.1, the input visual representation $\mathbf{V}$ has a shape of 512×60.

Audio representation: Short time Fourier transform (STFT) is firstly performed on the raw wave. Since the videos are at 25fps, to make the the audio aligned with the visual representation, the window size and hop length of STFT are set to 40ms and 10ms. With that STFT setup, there will be 4 frames of audio features aligned with their corresponding video frame. Audio representations are clipped or padded to a fixed length of 240. The audio is sampled at 16kHz, so the frequency resolution of the complex spectrogram is 321. Magnitude spectrogram has the same size as the complex spectrogram, which is 321×240. For phase spectrogram, we use $sine$ and $cosine$ value to present phase information, so the size is 642×240.

Synthetic Audio: To generate noisy audios, 2 utterances are randomly picked from the same dataset fold and mixed together.

4.3. Network Configuration

Structure: Details of $ResNets$ mentioned in section 2 and 3 are listed in Table 1. In order to reduce the number of parameters of the model, the filter number of all convolution layers is set to 1024, instead of 1536 in [13]. Phase sub-network is implemented following the recipe provided in [13]. Except the 3D visual ResNet front end

[3] Since we pad and clip the visual representation inputs to fixed length t, in our implementation, the weights of matrix $\mathbf{W}_A$ and $\mathbf{W}_B$ are shared. When computing matrix $\mathbf{Att}$, the padding positions are masked.

[4] Thanks for the open source code provided by Petridis[11], the project can be found at: https://sites.google.com/view/audiovisual-speech-recognition

Table 1. Network structure: **I**: The order of convolution layers in the ResNet; **C**: Number of convolution channels; **K**: Kernel size; **P**: Padding size; **S**: Convolution stride, $\frac{1}{2}$ for transposed convolutions; **RB**: Whether has a residual connection and batch normalization layer; **A**: The activation function.

Network	I	C	K	P	S	RB	A
$ResNet_V$	0	1024	5	2	1	×	None
	1-10	1024	5	2	1	✓	ReLU
$ResNet_M$	0	1024	5	2	1	×	None
	1	1024	5	2	1	✓	ReLU
	2	1024	5	2	2	✓	ReLU
	3	1024	5	2	1	✓	ReLU
	4	1024	5	2	2	✓	ReLU
	5	1024	5	2	1	✓	ReLU
$ResNet_F$	0	1024	5	2	1	×	None
	1-3	1024	5	2	1	✓	ReLU
	4	1024	5	2	$\frac{1}{2}$	✓	ReLU
	5-11	1024	5	2	1	✓	ReLU
	12	1024	5	2	$\frac{1}{2}$	✓	ReLU
	13-15	1024	5	2	1	✓	ReLU
	16	321	5	2	1	×	Sigmoid
$ResNet_P$	0	1024	5	2	1	×	None
	1-6	1024	5	2	1	✓	ReLU
	7	642	5	2	1	×	None

mentioned in section 4.2, all other main models are implemented with MXNet [23].

Network training. All magnitude sub-networks are trained with the same procedure. The networks are trained through two steps. Firstly, the initial learning rate is set to 10^{-3}. The learning rate is then reduced by the factor 0.7 for every 3 epochs. The Adam optimizer is used with the weight decay 10^{-5}. The gradient clipping is set to 10.0. Secondly, after the convergence, best model parameters in the validation set is chosen to optimize again. In the second optimizing procedure, the initial learning rate is set to 10^{-4}, and the weight decay is set to 10^{-6}. Then, after the convergence, the best model in the validation set is used for evaluation on the testing set. We do not make modification on the phase sub-network proposed in [13]. All magnitude sub-networks share a same phase sub-network, which is trained with the baseline magnitude sub-network. 4 GTX-1080Ti GPUs are used for data parallel training, and the mini batch size is set to 160.

4.4. Results and Analyze

LRS2 Dataset: Table 2 lists the results evaluated in LRS2 testing set, and different phases are used for the separated audio generation. The evaluation protocol includes the signal-to-distortion ratio score (SDR) [26] and the perceptual evaluation of speech quality score (PESQ) [27]. We first built the baseline by ourselves, and the system can get almost the same performance as that in [13]. Then the proposed architecture is constructed. Compared to the baseline, the 2-visual-stream model shows remarkable improvement in both SDR and PESQ score, and the designed attention mechanism can bring additional improvement. The system built with the newly proposed method is consistently better than the tradition method on all the conditions.

Control experiment: Results in Table 2 show the performance boost on the proposed attention based model. Considering that the model has extra parameters on the weight matrix in $\mathbf{W}_A$, $\mathbf{W}_B$, in order to do more fair comparison, we also constructed a model with the same structure but with a pseudo attention module. It has the same structure as the proposed attention based model, except that the attention weight matrix $\mathbf{Att}$ is sampled from a Gaussian noise, and the resluts are shown as the 4th line of Table 2. It shows that the pseudo attention module gets no additional improvement, which further demonstrates the effectiveness of our proposed attention mechanism.

Table 2. Performance evaluation on LRS2 dataset. **GT**: ground truth phase; **PR**: predicted phase; **MX**: noisy phase; **SDR**:signal to distortion ration, higher is better; **PESQ**: perceptual evaluation of speech quality, varies from -0.5 to 4.5, higher is better; * is the results copied from [13]. It is noted that their baseline model has a larger number of parameters than ours, with 1536 filters in every convolution layer, while our model only has 1024 filters.

Methods	Phase					
	GT		PR		MX	
	SDR	PESQ	SDR	PESQ	SDR	PESQ
1-Visual-Stream*	15.7	3.41	11.8	3.08	10.5	3.02
1-Visual-Stream	15.8	3.35	11.6	3.00	10.7	2.96
2-Visual-Stream	16.8	3.49	12.1	3.11	11.3	3.07
+ Pseudo Att	16.7	3.49	12.2	3.12	11.3	3.07
+ Att	**17.5**	**3.58**	**12.6**	**3.18**	**11.5**	**3.14**

Table 3. Performance evaluation on VoxCeleb2 dataset. **GT**: ground truth phase; **PR**: predicted phase; **MX**: noisy phase; **SDR**:signal to distortion ration, higher is better; **PESQ**: perceptual evaluation of speech quality, varies from -0.5 to 4.5, higher is better. The models trained on LRS2 are tested on VoxCeleb2 dataset directly.

Methods	Phase					
	GT		PR		MX	
	SDR	PESQ	SDR	PESQ	SDR	PESQ
1-Visual-Stream	7.0	2.40	2.9	2.20	2.8	2.18
2-Visual-Stream	9.8	2.74	5.9	2.49	5.5	2.47
+ Att	**10.7**	**2.86**	**6.7**	**2.57**	**6.1**	**2.56**

VoxCeleb2 Dataset: To evaluate the generalization of our proposed new model, we evaluate our models trained on LRS2 on the VoxCeleb2 dataset directly. The VoxCeleb2 dataset is collected from YouTube, while LRS2 is collected from BBC television. There is a mismatch between these two corpus. Most samples from VoxCeleb2 has lower video quality than those from LRS2. Moreover, LRS2 only consists of English speaker while VoxCeleb2 contains much more languages. The results are illustrated in Table 3. It is observed that the overall performance on VoxCeleb2 is worse than that on LRS2 due to the more challenging data. The proposed method can still obtain significant and consistent improvement compared to the conventional method on this generalization testing on VoxCeleb2.

5. CONCLUSION

In this paper, we explore a better strategy for audio-visual speech separation. Compared to the traditional method only using one target visual stream, the proposed new model takes advantages of both the speakers' visual stream, and it shows a better performance. Furthermore, by introducing an attention mechanism into the two visual streams between the target and interfered speakers, an additional and consistent improvement has been observed.

6. REFERENCES

[1] E Colin 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] Josh H McDermott, "The cocktail party problem," *Current Biology*, vol. 19, no. 22, pp. R1024–R1027, 2009.

[3] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji 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.

[4] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R Hershey, "Single-channel multi-speaker separation using deep clustering," *Interspeech 2016*, pp. 545–549, 2016.

[5] Zhong-Qiu Wang, Jonathan Le Roux, and John R Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5.

[6] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper 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.

[7] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, Jesper Jensen, Morten Kolbaek, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1901–1913, 2017.

[8] Yanmin Qian, Xuankai Chang, and Dong Yu, "Single-channel multi-talker speech recognition with permutation invariant training," *Speech Communication*, vol. 104, pp. 1–11, 2018.

[9] Zhuo Chen, Takuya Yoshioka, Xiong Xiao, Linyu Li, Michael L Seltzer, and Yifan Gong, "Efficient integration of fixed beamformers and speech separation networks for multi-channel far-field speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5384–5388.

[10] Takuya Yoshioka, Hakan Erdogan, Zhuo Chen, Xiong Xiao, and Fil Alleva, "Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks," *Proc. Interspeech 2018*, pp. 3038–3042, 2018.

[11] Stavros Petridis, Themos Stafylakis, Pingehuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic, "End-to-end audiovisual speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6548–6552.

[12] Themos Stafylakis and Georgios Tzimiropoulos, "Combining residual networks with lstms for lipreading," *Proc. Interspeech 2017*, pp. 3652–3656, 2017.

[13] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "The conversation: Deep audio-visual speech enhancement," *Proc. Interspeech 2018*, pp. 3244–3248, 2018.

[14] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 112, 2018.

[15] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*, 2016.

[16] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[17] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[18] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al., "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[19] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[20] Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou, "Abcnn: Attention-based convolutional neural network for modeling sentence pairs," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 259–272, 2016.

[21] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[22] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[23] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems," *arXiv preprint arXiv:1512.01274*, 2015.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[25] Ashutosh Pandey and Deliang Wang, "On adversarial training and loss functions for speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5414–5418.

[26] Cédric Févotte, Rémi Gribonval, and Emmanuel Vincent, "Bss_eval toolbox user guide–revision 2.0," 2005.

[27] Antony W Rix, John G Beerends, Michael P Hollier, and Andries 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)*. IEEE, 2001, vol. 2, pp. 749–752.