INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Time-domain Transformer-based Audiovisual Speaker Separation

Vahid Ahmadi Kalkhorani[1], *Anurag Kumar*[2], *Ke Tan*[2], *Buye Xu*[2], *and DeLiang Wang*[1]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Meta Reality Labs, USA

{ahmadikalkhorani.1, wang.77} @osu.edu
{anuragkr90, tanke1116, xub} @meta.com

Abstract

In this study, we propose a transformer-based architecture for talker-independent audiovisual speaker separation in the time-domain. Inputs to the proposed architecture are the noisy mixtures of multiple talkers and their corresponding cropped faces. Using a cross-attention mechanism, these two streams are fused together. The fusion layer is followed by a masking net that estimates one mask per talker and multiplies the mixed feature matrix by these masks to separate speaker features. Finally, the separated features are converted to the time domain at the decoder layer. Moreover, we propose a novel training strategy to increase the role of the video stream which starts with a relatively noisy condition and gradually increases audio stream quality during training. Experimental results demonstrate that the proposed method outperforms existing techniques according to multiple metrics on several commonly used audiovisual datasets.

Index Terms: audiovisual speaker separation, multi-modal speech processing, attentive audiovisual fusion.

1. Introduction

In real environments, speech usually occurs simultaneously with acoustic interference. The interference can be non-speech background noise and/or the speech signal from competing speakers. Joint audiovisual speech analysis plays a crucial role in human speech communication system and helps us focus attention on a specific talker and alleviate the effect of acoustic background noise. The integration of audio and visual information provides a more complete representation of speech and can improve human ability in understanding speech by as much as 40% [1] in noisy environments. But how to effectively leverage the information from both modalities remains a challenging task due to the intricate relationships between these modalities.

Recently, the use of deep neural networks (DNNs) for audiovisual speaker separation and speech enhancement has gained significant popularity and success [2]. The majority of these models use the time-frequency (T-F) representation of the audio signal [3, 4] and estimate T-F magnitude or magnitude-phase masks to separate speakers. Other studies perform this task in the time domain [5,6]. In terms of the video stream, various forms of visual data have been explored to analyze speech signal. These include single-frame images [7], lip area images and motion [8], and full-face videos [9]. While the lip area contains the most pertinent information for speech processing, recent studies [3, 10] demonstrate that using the entire face as input can be advantageous, particularly when the lip area is obstructed or when the talkers move their heads.

Most of the proposed techniques in the literature employ a convolutional neural network (CNN) followed by a set of feed-forward networks (FFNs) for audiovisual speech processing. These models process acoustic and visual data streams separately and employ a long-short-term memory (LSTM) netowrk to fuse features from these two modalities [2].

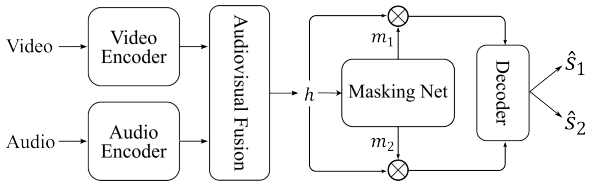

Figure 1: *The proposed model architecture for audiovisual speaker separation*

More recently, researchers have explored the use of attention-based mechanisms in audiovisual speaker separation (AVSS) [11] and speech enhancement (AVSE) [12]. Transformer models are designed to handle sequential data and can capture long-range dependencies between the input elements. This makes them well-suited for AVSS, where both the audio and visual signals are sequential in nature and exhibit dependencies between different time steps. Transformer-based models offer several advantages for AVSS. First, these models have the capability to identify long-term correlations between audio and visual inputs, which often exist in the AVSS domain. Secondly, the parallel processing nature of transformer models makes them well suited for AVSS applications that require real-time processing. Finally, transformer models have demonstrated strong performance in a range of sequential data processing tasks, including speaker separation [13] and speech enhancement [14], making them a promising approach for AVSS as well.

One of the challenges of joint audiovisual processing is the fact that the audio modality tends to dominate [15–17]. It is perceptually demonstrated that the contribution of visual cues to speech intelligibility is minor in relatively clean acoustic situations [1]. To maximize the contribution of the visual modality, Chung *et al.* [15] passed the data to their model in three modes: audio-only, video-only, and audio-video. Afouras *et al.* [16] added babble noise at 0dB SNR to the audio stream with a probability of 25% during training. Wei *et al.* [17] proposed an audiovisual fusion network that incorporates audio-video synchrony for audio representation and to improve video utilization.

In this work, we propose a time-domain audiovisual separation network based on dual-path attention architecture. The proposed model architecture for audiovisual speaker separation is shown in Fig. 1. For input, the architecture gets the noisy multi-talker mixtures from the audio modality and cropped faces of

10.21437/Interspeech.2023-2098

talkers from the visual modality. These two input streams are encoded separately and then passed to the attentive fusion layer. The fusion layer combines the encoded audio and video streams employing a cross-attentive mechanism. The masking net then separates the speaker features by estimating individual masks per speaker and applying them to the mixed feature matrix. Finally, the separated features are converted to the time domain at the decoder layer consisting of a transposed convolution layer. Additionally, we propose a training strategy to increase video utilization during the training by starting the training stage with a relatively noisy condition and increasing the audio stream quality gradually. Systematic evaluations and comparisons show that the proposed approach outperforms previous methods in multiple audiovisual speaker separation tasks.

2. Audiovisual Speaker Separation Mechanism

2.1. Problem Formulation

Given a video clip with an audio stream of N speakers talking $\sum_i^N s_i$ and a background noise n the noisy speech mixture in the time-domain is defined as

$$y(t) = \sum_{i=1}^{N} s_i(t) + n(t) \tag{1}$$

Audiovisual speaker separation aims at separating speaker speeches and suppressing the background noise using both audio and visual streams. We can formulate the audiovisual speaker separation using a deep neural network (DNN) as

$$\{\hat{s}_i | i \in [1, N]\} = f_\theta(y, V) \tag{2}$$

where V is the video streams and f_θ denotes DNN network parameterized by θ. Note that when $N = 1$, this task is converted to speech enhancement, its goal being to remove the background noise n from speech signal s_1.

2.2. Audiovisual speaker separation model

The initial step in the proposed method involves feeding the audio signal and video frames into separate encoders. The encoders are designed to transform the input data into a higher-dimensional feature space. The encoded video is then upsampled to ensure the number of frames in the audio and visual features match. The encoded audio and video are combined in the audiovisual fusion layer, producing a fused feature represented by the variable h. The fused feature is then passed to the MaskNet, which is responsible for estimating the masks. The estimated masks will be multiplied by the fused feature matrix enabling the separation of speaker features. The resulting separated features are then converted back to the time domain using a 1D transposed convolution layer. Since the output of the model is the estimated waveform of clean signals, we train the model via utterance-level Permutation Invariant Training (uPIT) and Scale-Invariant Signal-to-Noise Ratio (SI-SNR) loss function.

2.3. Audio encoder

For the audio encoder, we use a 1D convolution layer. This encoder utilizes a kernel size of k and a stride of $k/2$ to process the mixed noisy signal $y \in \mathbb{R}^{T_a}$ in the time domain. The audio encoder's objective is to extract features, denoted as $h_a \in \mathbb{R}^{F_a \times T'_a}$ from noisy audio signal.

2.4. Video encoder

For the video encoder as shown in Fig. 2(a), we leverage a model that is based on the Lip2Wav architecture [18]. This block takes cropped speaker faces presented in the mixed audio signal as input $V \in \mathbb{R}^{T_v \times 3 \times 160 \times 160}$. This model employs a sequence of blocks with 3D convolutions, normalization, and Rectified Linear Unit (ReLU) activation function, followed by a bidirectional LSTM. The encoded video matrix of all speakers is concatenated on feature dimension and is denoted as $h'_v \in \mathbb{R}^{F_v \times T_v}$. Then, the encoded video is up-sampled in the time dimension using the nearest method to match the number of frames in the encoded audio. The output of this block is denoted as $h_v \in \mathbb{R}^{F_v \times T'_a}$.

2.5. Attentive audiovisual fusion

In the audiovisual fusion block, we use an attentive fusion network similar to [17]. The structure of the fusion block is shown in Fig. 2(b). This block involves feeding encoded audio and video features into separate multi-head-attention (MHA) [19] blocks. Following this, we designate the video (audio) features as K and V vectors, and audio (video) features serve as the query vector Q for the next *MHA* on the audio (video) side.

$$f'_a = MultiHead(Q_a, K_v, V_v) \tag{3}$$

$$f'_v = MultiHead(Q_v, K_a, V_a) \tag{4}$$

These blocks are subsequently followed by a layer normalization and a FFN layer. To process the fused features, the output of these two streams is concatenated and passed through a final round of MHA attention and FFN blocks. The goal of this process is to extract and map data in the fused feature domain. As demonstrated by [17], this technique effectively lessens the disparity between audio and visual modalities. The fused audiovisual output matrix $h \in \mathbb{R}^{F_a \times T'_a}$ is then passed to the masking net.

2.6. Masking Net

For the masking net, we utilize a dual-path attention network as proposed in [13] to generate a mask matrix for each talker. The configuration of this network is illustrated in Figure 2(c). First, we normalize the fused audiovisual features and pass the result to a linear layer. The output is then segmented into chunks of time with 50% overlap between each pair of consecutive chunks. Subsequently, we apply a dual-path attention layer, similar to the SepFromer block proposed in [13]. This block is composed of two transformer blocks, which share a similar structure, with the only distinction being that the first block processes the time sequence, while the second one processes the feature sequence. As shown in 2(c), each transformer block is comprised of a layer normalization and a self-attention block followed by another set of layer normalization and FFN layer. To optimize training efficiency, residual connections are incorporated enabling gradients to flow through the layers. Then, the output is fed to an overlap-and-add-block and a linear layer, as described in [20]. Finally, a linear layer is utilized in combination with a ReLU function to estimate one mask matrix per talker.

2.7. SNR scheduler

In order to increase the impact of visual data on the model's performance, we implement a technique named SNR scheduler. In

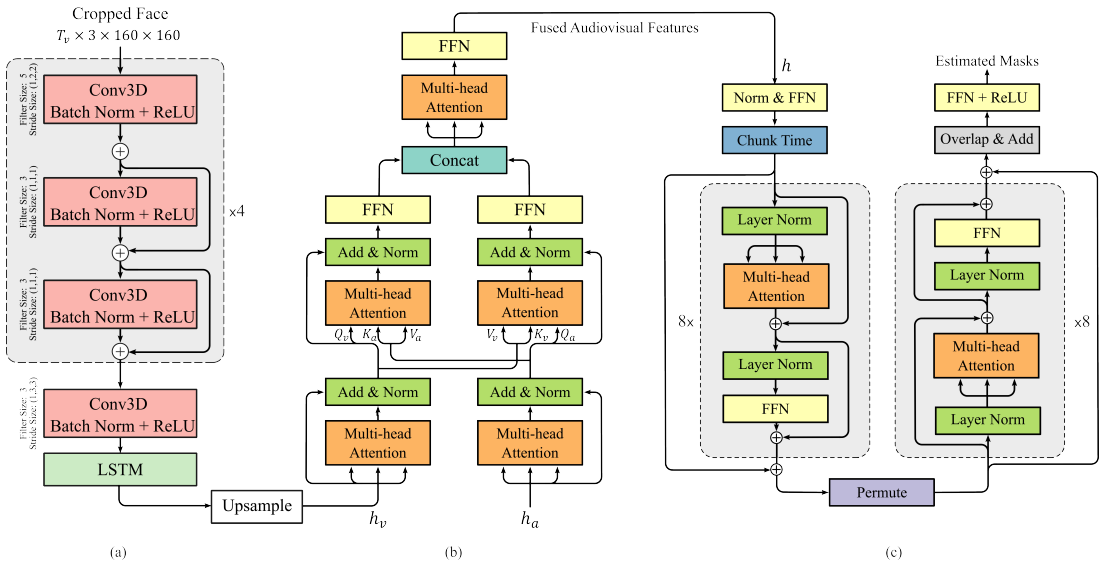

Figure 2: *Illustration of the proposed audiovisual speaker separation architecture (a) Video encoder, (b) Fusion block, (c) Masking block*

this approach, we initiate the training process with high levels of background noise which we gradually decrease over time. The information carried by the visual stream remains intact as it is not impacted by background noise or reverberation. As a result, the high noise levels in this approach prevent the audio features from dominating the model input, allowing the visual modality to play a more prominent role in the early stages of training. As training progresses, we gradually decrease the background noise energy level, allowing the audio features to become dominant once again.

3. Experimental Setup

3.1. Datasets

We use AVSpeech dataset [9] for the training. This dataset consists of over 4,700 hours of 3- to 10-second-long video segments gathered from 290,000 YouTube videos. The collection includes speakers from a wide range of acoustic conditions, including almost 150,000 individuals who speak a variety of languages and have varied vocabularies. The AVSpeech dataset samples may contain background noise and reverberation, which may provide difficulties during the training process due to their uncleanliness. To address this limitation, we use a subset of the AVSpeech dataset. To identify relatively clean samples, we employ CMGAN [21], to enhance the signals. Then, we measure the Signal-to-Noise Ratio (SNR) value of the original audio relative to the enhanced signal and filter out samples with an SNR value lower than 15dB, leading to around 700k samples. In the training stage, we randomly choose two 3-second samples in the training stage from this subset. For the test and validation sets, we randomly separate AVSpeech test dataset into two subsets and choose 3000 pairs from each for validation and testing. We combine the selected speech signal pairs using a signal-to-inference ratio (SIR) drawn from (-2.5, 2.5) dB. For the background noise, we randomly choose 3-second signals from AudioSet [22] noise collection, which consists of nearly 1.7 million 10-second segments comprising 526 distinct forms of noise. We then select a signal-to-noise ratio (SNR) value between (-2.5, 2.5)dB and add the chosen noise with this SNR to the mixed-talkers signal.

3.2. Pre-processing and augmentation

In our study, we utilize an audio sampling rate of 8kHz and a video frame rate of 25 fps. During the training process, we adjust the maximum of the time-domain mixture to one and apply the same scaling factor to each of the clean sources.

We crop speaker faces from the video stream employing MediaPipe [23] library with a margin of 10 pixels. Then, we resize each cropped image to 160×160 *px*. We use RGB images and scale them to range [-1, 1] range. In each epoch, we sample a 3-second segment from the synthetic mixtures. This corresponds to 3×8000 audio samples and 3×25 video frames.

We also employ Dynamic mixing (DM) data augmentation approach, as outlined in [13]. This technique involves the generation of new mixtures on-the-fly using random single-talker sources. Additionally, we augment the video stream by shifting the video by (-5,+5) frames in time. This augmentation technique aims to improve the model's ability to handle scenarios in which the audio and video streams may not be perfectly synchronized.

3.3. Experiments Settings

We use 256 convolutional filters with a kernel size of 16 samples and a stride factor of 8 samples for the audio encoder. The kernel size and stride values of the video encoder are shown in Fig. 2. We employ a bidirectional LSTM at the last layer of the video encoder with a hidden state size of 384. After concatenating two speakers encoded features, this results in a feature vector of size 384×2.

In the training phase, we utilize the Adam optimizer [24] with a learning rate of 1.5×10^{-4}. When the validation performance does not improve for five consecutive epochs, we reduce the learning rate by half. We use a batch size of 2 and train the model on 8 nodes with 4 GPUs on each node (32 GPUs in total). To speed up training and minimize memory usage, we utilize automatic mixed precision. We employ SI-SNR via utterance-level permutation invariant loss [25] to train the model.

Table 1: *Audiovisual speaker separation results on the AVSpeech [3] dataset with noise signals from AudioSet [22]. SI-SDRi: scale-invariant signal-to-distortion ratio improvement. Higher is better.*

	SNR sch.	Fusion	SI-SDRi
Audio Only [13]	✗	-	11.22
Ephrat *et al.* [3]	✗	LSTM	10.71
Proposed	✗	LSTM	11.51
	✓	LSTM	13.15
	✓	Attentive	**13.35**

3.4. Experiments Results

3.4.1. Audiovisual fusion mechanism and SNR scheduler

We present the result of speaker separation on AVSpeech [3] test dataset in Table 1. In this table, we compare the results of our proposed audiovisual model, which features a fusion block and SNR scheduler, to two other models: an audio-only model with a structure similar to ours [13], and an audiovisual model with an LSTM block as the fusion layer [3].

To create the test mixture, we utilize the methodology outlined in the paper by [3]. This involves choosing 2000 pairs of random audio clips, with i^{th} audio signal pair denoted as s_1^i and s_2^i. Additionally, we select a noise signal, denoted as n^i, from the AudioSet noise dataset. We then combine the audio clips and noise signal by adding them together with a weighting factor of 0.3, resulting in the i^{th} mixed audio signal $s_{mix}^i = s_1^i + s_2^i + 0.3n^i$.

To evaluate the impact of the fusion block and SNR scheduler, we present the results of our proposed structure with three different configurations. As shown in Table 1, incorporating the SNR scheduler has a significant impact on model performance. Moreover, the performance of the audio-only model is comparable with the audiovisual model when it is trained without employing the SNR scheduler technique. This is showing that in this case, the model is not able to leverage visual modality. However, when we employ the SNR scheduler technique the performance improves significantly. To examine the fusion block's effect on the model performance, we present the outcomes of the model using an LSTM and attentive fusion block in Table 1. We observe that the attentive fusion block, combined with the SNR scheduler, further improves the model performance.

3.4.2. Comparison with prior works

We present a comparison of our results with those of state-of-the-art models on three audiovisual datasets, which can be found in Table 2. Mandarin dataset [26] comprises 320 video recordings of a Mandarin speaker uttering sentences, with each sentence consisting of 10 Chinese characters. Each utterance was recorded in a noise-free environment, with the speaker directly facing the camera. LRS3 [27] test set includes videos from TED and TEDx talks of 451 speakers with a total duration of 1 hour. VoxCeleb2 [28] test set contains in-the-wild video clips of celebrities with 118 distinct identities. This dataset includes several challenging scenarios such as varied video quality, low lighting, and recordings made from a side view.

Our approach is similar to that of [3, 8], where we train the model on the AVSpeech dataset and test it on the other datasets to assess both performance and cross-corpus generalization capability. To evaluate the results, we use Perceptual Evaluation of Speech Quality (PESQ) [29], Signal-to-Distortion Ratio, and Short-Time Objective Intelligibility (STOI) [30] metrics. Result of VisualVoice [8] on LRS3 dataset in this table is from [31].

Table 2(a) presents the results of our model on the Mandarin dataset [26]. It is important to note that this dataset comprises single-talker noisy mixtures and is utilized as a baseline for evaluating audiovisual models on the speech enhancement task. To assess our model's performance in this situation, we train the model using one-talker signals mixed with background noise.

As presented in Table 2, our proposed model and training technique consistently outperforms the other methods on speech enhancement and speaker separation tasks.

Table 2: *Comparison with privoius works*

(a) Mandarin (Enhancement) [26]

	PESQ	SDR
Noisy	2.09	
Gabbay *et al.* [32]	2.25	-
Hou *et al.* [26]	2.42	2.8
Ephrat *et al.* [3]	2.50	6.1
VisualVoice [8]	2.51	6.69
Proposed	**2.67**	**7.42**

(b) LRS3 [27]

	PESQ	SDR
Noisy	1.30	
Lee *et al.* [9]	2.01	9.78
VisualVoice [8]	2.41	-
Proposed	**2.81**	**10.02**

(c) VoxCeleb2 [28]

	PESQ	SDR	STOI
Noisy	2.13	0.05	0.70
VoVit [11]	-	10.03	0.87
VisualVoice [8]	2.83	10.2	0.87
Proposed	**2.94**	**11.52**	**0.88**

4. Conclusion

This study focuses on audiovisual speaker separation in noisy environments. We have proposed a time-domain audiovisual integration model for single-channel speaker separation. The proposed method incorporates a MaskNet architecture with dual-path attention and a 3D Encoder for audio and visual processing. We have also presented a new training strategy called SNR scheduler to increase visual utilization in audiovisual integration. The resulting model achieves better separation results than recent baselines on several commonly used audiovisual datasets. Future work will include the development of causal speech enhancement and speaker separation and expanding the proposed architecture to incorporate multi-channel acoustic features.

5. Acknowledgements

This research was supported in part by a research contract from Meta Reality Labs and Illinois Supercomputer Center (NSF ACI-1928147).

6. References

[1] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[2] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, 2021.

[3] 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.

[4] A. Arriandiaga, G. Morrone, L. Pasa, L. Badino, and C. Bartolozzi, "Audio-visual target speaker enhancement on multi-talker environment using event-driven cameras," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2021, pp. 1–5.

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

[6] E. Ideli, B. Sharpe, I. V. Bajić, and R. G. Vaughan, "Visually assisted time-domain speech enhancement," in *2019 IEEE global conference on signal and information processing (GlobalSIP)*. IEEE, 2019, pp. 1–5.

[7] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang, "Facefilter: Audio-visual speech separation using still images," *arXiv preprint arXiv:2005.07074*, 2020.

[8] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pp. 15 490–15 500.

[9] 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.

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

[11] J. F. Montesinos, V. S. Kadandale, and G. Haro, "Vovit: Low latency graph-based audio-visual voice separation transformer," *arXiv preprint arXiv:2203.04099*, 2022.

[12] R. Mira, B. Xu, J. Donley, A. Kumar, S. Petridis, V. K. Ithapu, and M. Pantic, "La-voce: Low-snr audio-visual speech enhancement using neural vocoders," *arXiv preprint arXiv:2211.10999*, 2022.

[13] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25.

[14] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1598–1607, 2020.

[15] J. Son Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6447–6456.

[16] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 8717–8727, 2018.

[17] L. Wei, J. Zhang, J. Hou, and L. Dai, "Attentive fusion enhanced audio-visual encoding for transformer based robust speech recognition," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2020, pp. 638–643.

[18] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "Learning individual speaking styles for accurate lip to speech synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 796–13 805.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.

[21] R. Cao, S. Abdulatif, and B. Yang, "Cmgan: Conformer-based metric gan for speech enhancement," *arXiv preprint arXiv:2203.15149*, 2022.

[22] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[23] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

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

[25] 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.

[26] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

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

[28] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[29] 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.

[30] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[31] A. Rahimi, T. Afouras, and A. Zisserman, "Reading to listen at the cocktail party: Multi-modal speech separation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 493–10 502.

[32] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement using noise-invariant training," *arXiv preprint arXiv:1711.08789*, 2017.