ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446297

AUDIOVISUAL SPEAKER SEPARATION WITH FULL- AND SUB-BAND MODELING IN THE TIME-FREQUENCY DOMAIN

Vahid Ahmadi Kalkhorani[1]*, Anurag Kumar*[2]*, Ke Tan*[2]*, Buye Xu*[2]*, and DeLiang Wang*[1,3]

[1]Department of Computer Science and Engineering, Ohio State University, USA
[2]Meta Reality Labs,USA, [3] Center for Cognitive and Brain Sciences, The Ohio State University, USA

ABSTRACT

We introduce a new deep learning model for talker-independent audiovisual speaker separation in noisy conditions in the time-frequency domain. The inputs to the model include noisy multi-talker mixtures and the corresponding cropped face images. Our approach incorporates cross-attention audiovisual fusion, effectively merging audio and visual features and enabling seamless information interchange between auditory and visual modalities. These fused features drive a separator module, which separates the acoustic features of individual speakers. The separator module is based on the recently proposed TF-Gridnet, which comprises an intra-frame full-band component, a sub-band temporal module that captures frequency-specific temporal dependencies, and a cross-attention module dedicated to extracting long-term fused audiovisual features. To encourage the utilization of visual streams during training, we employ a Signal-to-Noise Ratio (SNR) scheduler. Experimental results demonstrate that the proposed model advances the state-of-the-art speaker separation performance in several audiovisual benchmark datasets.

Index Terms— audiovisual speaker separation, multimodal speech processing, attentive audiovisual fusion

1. INTRODUCTION

In human speech communication, the presence of acoustic interference, such as background noise or competing speakers, presents challenges for speech understanding. In such environments, the availability of visual information can mitigate the impact of background interference. Integrating complementary audio and visual information is shown to result in improvements in speech comprehension, especially in highly noisy environments [1].

Multi-talker speaker separation is traditionally tackled using statistical methods [2, 3, 4]. In recent years, deep neural networks (DNNs) have gained popularity in audiovisual speaker separation (AVSS) and audiovisual speech enhancement (AVSE) [5]. DNN models for speaker separation typically consist of three modules: 1) an acoustic feature extractor, 2) a visual feature extractor, and 3) a fusion and separator module. In the early DNN models, the primary acoustic input feature was the magnitude spectrogram of the noisy speech in the time-frequency domain [6]. Recent research has expanded acoustic features in audiovisual (AV) systems. These features include both the magnitude spectrogram and the respective phase [7], the real and imaginary parts of the complex spectrogram [8], or even the raw waveform [9].

Researchers have explored various visual data types for AVSS, including single-frame images [10, 11], lip area images, and lip motion, as well as facial landmark points [12] and full-face videos [13]. Employing the entire face as the input has shown advantages, especially in challenging situations where the lip area is obstructed or when speakers are in motion [14]. In AVSS, a prevalent model architecture combines convolutional neural networks (CNNs) with multilayer perceptrons (MLPs) [15, 16, 17] to process acoustic and visual data independently, extracting distinct features. These features are then fused using long short-term memory (LSTM) to capture temporal correlations inherent in audio and visual signals. Recent research has incorporated transformer-based mechanisms in AVSS and AVSE [18]. Transformer-based models, originally designed for natural language processing, show promise in AVSS by capturing long-range dependencies between audio and visual inputs for better integration of information across time steps.

An inherent challenge in AVSS is the tendency for the audio modality to dominate the visual modality [19, 20]. Perceptual research demonstrates that in relatively clean acoustic environments, the contribution of visual cues to speech intelligibility is relatively minor; however, the contribution becomes significant in very noisy acoustic conditions [1]. To maximize the utilization of the visual modality, various approaches have been explored in the literature [19, 20]. Although these studies have investigated the AV integration for speaker separation, how to harness information from both modalities for high-quality separation remains unclear.

In this study, we present a novel AVSS network called AV-GridNet, based on the recently proposed TF-Gridnet which performs complex spectral mapping for speech separation in the time-frequency (T-F) domain [21]. We utilize a cross-

This research was supported in part by a research contract from Meta Reality Labs and Ohio Supercomputer Center

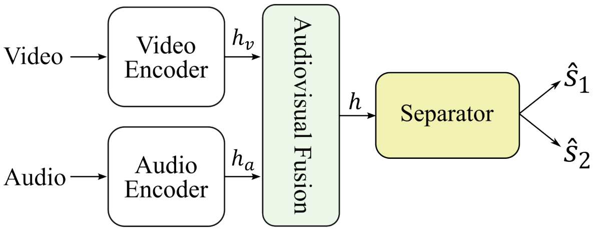

Fig. 1: Overall architecture of the proposed audiovisual model for speaker separation

attention AV fusion module for the exchange of information between audio and visual modalities, allowing the model to capture complementary cues from these modalities. To further enhance the model's ability to leverage visual information for speaker separation, we employ a signal-to-noise ratio (SNR) scheduler to emphasize the contribution of the visual stream during the training.

Training our proposed model on AVSpeech [13] dataset, we assess its performance on four benchmark audiovisual datasets. The results demonstrate state-of-the-art performance in audiovisual speaker separation, showcasing the effectiveness of our proposed model and training algorithm.

The rest of the paper is organized as follows: Section 2 describes the building blocks of the proposed AV model. Experimental settings are given in Section 3, and results and comparisons are presented in Section 4. Finally, the concluding remarks are given in Section 5.

2. MODEL DESCRIPTION

The overall architecture of our proposed model is illustrated in Fig. 1. The audio encoder takes the mixed-signal in short-time Fourier transform (STFT) domain and extracts acoustics features employing a 2D convolutional layer (Conv2D). The video component works with 3-channel (RGB) cropped faces of the speakers in the scene. A visual encoder is utilized to extract visual features from the cropped faces of all speakers in each frame of the video clip. Once we obtain the processed audio and visual features, we fuse them using a cross-attention module. This fusion step combines information from both modalities to capture their mutual dependencies. The fused audiovisual features are then processed by a sequence of *B* separator blocks, responsible for separating the feature matrix of each speaker. The separated feature matrices are finally converted back to the time domain using a one-dimensional deconvolution (Deconv1D) layer.

2.1. Audio encoder

Once the mixed-noisy signal has been transformed into the T-F domain, we employ a Conv2D with a 3×3 kernel to process the T-F representations. This convolutional operation helps capture local patterns and dependencies within the T-F units. Following the Conv2D operation, we apply global layer normalization (gLN) [22] to normalize the activations across the T-F units. After normalization, we obtain a *D*-dimensional embedding for each T-F unit. These embeddings collectively form an acoustic feature tensor $h_a \in \mathbb{R}^{D\times T_a\times F}$, where D represents the embedding dimension, T_a represents the number of time frames, and F represents the number of frequency bins.

2.2. Video encoder

To extract visual features, first, we extract the speaker faces in each video frame and resize them to $160 \times 160\ px$. This process enables us to generate 25 face thumbnails per speaker per second, assuming that the video is recorded at a frame rate of 25 frames per second (FPS). Then, we use the Inception ResNet V1 model [1] to encode face images and obtain a visual feature matrix h_v. We modify the last layer of the Inception model to match the dimension of the audio features F. Although we use a pre-trained model, we do not freeze its parameters and allow them to be trained for the task of speaker separation. This enables the model to learn and adapt visual representations that are most suitable for separating the speakers in the given context.

2.3. Attentive audiovisual fusion

We adopt a cross-modal attentive fusion network [20, 23]. As shown in Fig. 2(a), this block leverages multi-head attention (MHA) blocks to effectively combine information from encoded audio and video streams. First, we feed the encoded audio and video features into separate MHA blocks. Within these blocks, we designate the video (audio) features as K and V vectors, while the audio (video) features serve as the query vector Q for the subsequent MHA operation on the audio (video) side. Then, we use layer normalization (LN) and a feed-forward layer to process the output of each modality. Finally, we concatenate the output from the audio and video streams. This concatenated output then goes through a final round of MHA and feed-forward network (FFN). The main objective of the last step is to effectively extract and map information in the fused feature space.

2.4. Separator Module

The separator module, Fig. 2(b), consists of four key components: the intra-frame full-band module Fig. 2(c), sub-band temporal module Fig. 2(d), cross-attention module, and a two-dimensional deconvolution (Deconv2D) Layer. The intra-frame full-band module uses a bidirectional long short-term memory (BLSTM) layer to extract the correlation between the local spectral information within each frame of the fused AV features. The sub-band temporal module also employs BLSTM to capture feature correlations over time. The

[1] https://github.com/timesler/facenet-pytorch

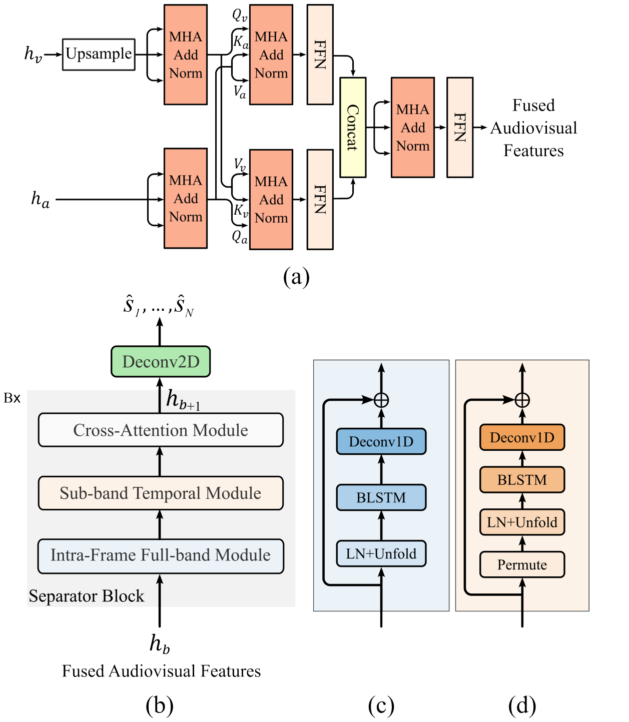

Fig. 2: Building blocks of proposed model. (a) Attentive AV fusion, (b) Separator module, (c) intra-frame full-band module, (d) sub-band temporal module

cross-attention module [21] is responsible for capturing long-range global information by using multi-head attention on the frame embeddings. Finally, the Deconv2D Layer reconstructs the separated audio features back into their original time-domain representation.

2.5. SNR Scheduler

To enhance the impact of the visual modality on the model's performance, we employ a technique known as SNR scheduler [23]. This method emphasizes the importance of the visual stream by manipulating background noise levels during the training. Initially, training begins with high background noise, deliberately creating a challenging environment where speech signals contend with heavy noise corruption. By introducing substantial noise, the visual modality gains more influence, enabling effective use of visual cues despite significant acoustic interference. As training progresses, we decrease background noise energy gradually until it matches the validation/test SNR bounds.

3. EXPERIMENTAL SETUP

3.1. Datasets and preprocessing

During training, we utilize the AVSpeech dataset [19], which aggregates video clips from diverse YouTube sources. This dataset comprises 4,700+ hours of 3 to 10-second segments, covering varying acoustic conditions, including noise and reverberation. To address the issue of uncleanliness in the AVSpeech dataset, we adopt a subset of the data comprising relatively clean samples. To identify these clean samples, we utilize a pre-trained audio-only speech enhancement model CMGAN [24]. To filter out the clean samples, we compare the original audio signals with the enhanced signals and assess SNR. Samples with an SNR value below 15dB are discarded, resulting in a refined subset of approximately 700k samples. During the training process, we randomly select two 3-second samples from this subset and combine them to create training pairs. For the background noise, we randomly choose 3-second signals from the noise collection provided by AudioSet [25]. AudioSet contains nearly 1.7 million 10-second segments, encompassing 526 distinct forms of noise. From this collection, we select a background noise signal and determine a SNR value between (-20, 20) dB. We add the chosen noise signal to the mixed-speakers signal with the specified SNR, resulting in a realistic audio mixture that includes speech from multiple speakers and background noise.

3.2. Experimental Settings

In our experimental setup, we utilize a STFT with a window length of 32 ms and a hop length of 8 ms. The analysis window employed is a square-root Hann window. To obtain complex STFT spectra, we employ a 256-point discrete Fourier transform (DFT), resulting in 129-dimensional spectra. For the self-attention mechanism, we employ a point-wise Conv2D operation with 4 output channels to generate key and query tensors. Each T-F unit is embedded in a 48-dimensional space, and the hidden units of BLSTMs are set to 192. In the cross-attention module, we employ 4 attention heads. To capture a comprehensive representation, we conduct all experiments using 6 separator blocks. To train the model, we employ scaled-estimate scale-invariant signal-to-distortion ratio (SE-SI-SDR) loss function [21]. To calculate this form of SI-SDR loss function, we scale the estimated signal to equalize its gain with that of the source signal.

4. EVALUATION RESULTS

In this section, we evaluate the performance of our model on four benchmark audiovisual datasets. Our approach is similar to previous works such as [13, 26], in which we train our model on the AVSpeech dataset and assess its effectiveness and generalization capability by evaluating its performance on other datasets.

The outcomes of our speaker separation experiments on the AVSpeech test dataset are displayed in Table 1(a). This table compares the results of our proposed audiovisual model with three other models: an audio-only model [27], an audiovisual model utilizing an LSTM block for fusion [13], and a time-domain transformer based audiovisual model [23]. To generate the test mixture, we follow the methodology outlined in [13]. To assess the impact of the SNR scheduler on the model performance, we present the results of our proposed model with two different configurations. As demonstrated in Table 1(a), the incorporation of the SNR scheduler significantly influences the performance of the model.

Table 1: Comparison with previous speaker separation works

(a) AVSpeech [13]

	SNR sch.	SI-SDRi
Audio Only [27]	✗	11.22
Ephrat *et al.*[13]	✗	10.71
Ahmadi *et al.*[23]	✓	13.15
Proposed	✗	13.24
Proposed	✓	**13.85**

(b) LRS3 [28]

	PESQ	STOI
Unprocessed	1.30	0.75
Lee *et al.*[14]	-	0.85
VisualVoice [26]	2.41	0.90
Rahimi *et al.*[29]	2.42	0.94
Proposed	**2.54**	**0.95**

(c) VoxCeleb2 [30]

	PESQ	SDR	STOI
Unprocessed	1.58	0.09	0.64
VoVit [12]	-	10.03	0.87
VisualVoice [26]	2.83	10.2	0.87
Ahmadi *et al.*[23]	2.94	11.54	0.88
Proposed	**3.05**	**12.37**	**0.90**

To evaluate the performance of our model on the LRS3 dataset [28], we follow the methodology described in [14], using the same set of video clips to generate 1320 test samples. Table 1(b) presents the results of our model on the LRS3 dataset. As depicted in the table, our model exhibits superior performance in terms of PESQ and STOI metrics compared to other methods.

Table 1(c) presents the results of our model on the VoxCeleb2 [30] dataset. The VoxCeleb2 dataset consists of in-the-wild video clips featuring 118 distinct celebrities. This dataset is particularly challenging due to its varied video quality, low lighting conditions, and recordings captured from different perspectives, such as a side view. We generate 3000 random samples without any background noise. As shown in the table, our proposed model achieves superior performance compared to other methods, as indicated by the higher scores in terms of PESQ and STOI metrics. This demonstrates the effectiveness of our approach in handling the challenges posed by the VoxCeleb2 dataset and highlights its potential for real-world audiovisual speaker separation tasks.

Fig. 3 presents a comprehensive performance comparison between our proposed audiovisual model and an audio-only (AO) model using the GRID dataset [31]. The evaluation was conducted under various mixed-SNR conditions, spanning from -20 dB to +20 dB. Notably, the results demonstrate the superior performance of our audiovisual model over the audio-only counterpart across this wide range of SNR values. The outcomes obtained from the comparison emphasize the significance of integrating both audio and visual information to achieve enhanced speech separation, especially in noisy environments. Furthermore, the results presented in Fig. 3 offer an insightful observation. As the background noise energy level decreases, the discrepancy between the audio-only model and the audiovisual model diminishes. This phenomenon highlights the growing contribution of the visual modality as the signal becomes increasingly noisy. The visual cues appear to play a more prominent role in disentangling speech from noise in challenging acoustic conditions, leading to the notable performance gains of the audiovisual model.

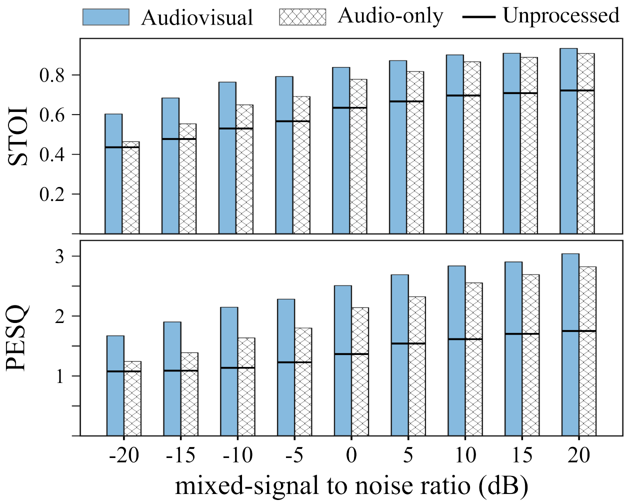

Fig. 3: Performance comparison of the proposed model and audio-only model on GRID dataset [31]

5. CONCLUSION

This study focuses on the problem of audiovisual speaker separation in noisy environments. We have proposed a novel time-frequency domain audiovisual model designed for single-channel speaker separation. To integrate audio and visual features, we employ a cross-attention fusion module. To enhance the utilization of visual cues in the audiovisual integration process, we introduce a training strategy called an SNR scheduler. This strategy dynamically adjusts SNR during training, in order to increase the contribution of visual information for improved robustness. Experimental results demonstrate that our proposed model outperforms recent baselines on various widely used audiovisual datasets. Future work will include the development of causal speaker separation and expanding the proposed architecture to incorporate multi-channel acoustic features.

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, pp. 212–215, 1954.

[2] I. Almajai and B. Milner, "Visually derived wiener filters for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 1642–1651, 2010.

[3] M. S. Khan, S. M. Naqvi, W. Wang, J. Chambers *et al.*, "Video-aided model-based source separation in real reverberant rooms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 1900–1912, 2013.

[4] Y. Liang, S. M. Naqvi, and J. A. Chambers, "Audio video based fast fixed-point independent vector analysis for multisource separation in a room environment," *EURASIP journal on Advances in Signal Processing*, vol. 2012, pp. 1–16, 2012.

[5] 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.

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

[7] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. INTERSPEECH*, 2018, pp. 3244–3248.

[8] B. İnan, M. Cernak, H. Grabner, H. P. Tukuljac, R. C. Pena, and B. Ricaud, "Evaluating audiovisual source separation in the context of video conferencing," in *Proc. INTERSPEECH*, 2019, pp. 4579–4583.

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

[10] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang, "FaceFilter: Audio-Visual Speech Separation Using Still Images," in *Proc. INTERSPEECH*, 2020, pp. 3481–3485.

[11] L. Qu, C. Weber, and S. Wermter, "Multimodal Target Speech Separation with Voice and Face References," in *Proc. INTERSPEECH*, 2020, pp. 1416–1420.

[12] 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.

[13] 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.

[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] R. Lu, Z. Duan, and C. Zhang, "Listen and look: Audio–visual matching assisted speech source separation," *IEEE Signal Processing Letters*, vol. 25, pp. 1315–1319, 2018.

[16] M. Gogate, A. Adeel, R. Marxer, J. Barker, and A. Hussain, "DNN Driven Speaker Independent Audio-Visual Mask Estimation for Speech Separation," in *Proc. INTERSPEECH*, 2018, pp. 2723–2727.

[17] S.-Y. Chuang, Y. Tsao, C.-C. Lo, and H.-M. Wang, "Lite Audio-Visual Speech Enhancement," in *Proc. INTERSPEECH*, 2020, pp. 1131–1135.

[18] D. Ivanko, D. Ryumin, and A. Karpov, "A review of recent advances on deep learning methods for audio-visual speech recognition," *Mathematics*, vol. 11, p. 2665, 2023.

[19] 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, pp. 8717–8727, 2018.

[20] 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.

[21] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "Tf-gridnet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[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, pp. 1256–1266, 2019.

[23] V. Ahmadi Kalkhorani, A. Kumar, K. Tan, B. Xu, and D. Wang, "Time-domain Transformer-based Audiovisual Speaker Separation," in *Proc. INTERSPEECH*, 2023, pp. 3472–3476.

[24] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based Metric GAN for Speech Enhancement," in *Proc. INTERSPEECH*, 2022, pp. 936–940.

[25] 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.

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

[27] 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.

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

[29] 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.

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

[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, pp. 2421–2424, 2006.