ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096893

ROBUST AUDIO-VISUAL ASR WITH UNIFIED CROSS-MODAL ATTENTION

Jiahong Li, Chenda Li, Yifei Wu, Yanmin Qian

MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
{li_jiahong,lichenda1996,yifei.wu,yanminqian}@sjtu.edu.cn

ABSTRACT

Audio-visual speech recognition (AVSR) takes advantage of noise-invariant visual information to improve the robustness of automatic speech recognition (ASR) systems. While previous works mainly focused on the clean condition, we believe the visual modality is more effective in noisy environments. The challenges arise from the difficulty of adaptive fusion of audio-visual information and the possible interferences inside the training data. In this paper, we present a new audio-visual speech recognition model with a unified cross-modal attention mechanism. In particular, the auxiliary visual evidence is combined with the acoustic feature along the temporal dimension in the unified space before the deep encoding network. This method provides a flexible cross-modal context and requires no forced alignment such that the model can learn to leverage the audio-visual information in relevant frames. In experiments, the proposed model is demonstrated to be robust to the potential absence of the visual modality or misalignment in audio-visual frames. On the large-scale audio-visual dataset LRS3, our new model further reduces the state-of-the-art WER for clean utterances and significantly improves the performance under noisy conditions.

Index Terms— audio-visual speech recognition, unified cross-modal attention, noise-robust, modality absence

1. INTRODUCTION

Automatic Speech Recognition (ASR) can be a difficult task when the speech signal gets distorted by noises. As in the conversation between humans, visual modality sometimes becomes necessary to understand speeches accurately since it's invariant to the presence of noise. Audio-Visual Speech Recognition (AVSR) is the task of generating text transcriptions from both the audio and the auxiliary visual evidence. Although many kinds of visual sources [1] can be useful for speech recognition, lip motion [2] [3] is considered the most related and beneficial evidence in the literature.

Quite a few works have demonstrated the complementary effect in AVSR by fusing the visual stream into mature ASR models. The transformer-based encoder-decoder structure [4] showed great performance with the audio-visual representations from separate encoders concatenated together for decoding. The hybrid CTC/Attention architecture [5] was also proven effective in AVSR, and it became the common architecture in researches afterward. Ma et al. [6] changed the transformer in encoders into the convolution-augmented transformer (conformer [7]), leading to the state-of-the-art performance. To make better use of the relatively weak visual information, researchers attempted to guide the models into learning the trade-off between modalities by gated structures [8] or global attention mechanism [9] [10]. Another challenge comes from the limited resources of labeled audio-visual data, requiring the model to be able to endure the asynchronisation issue between modalities. Several works [11] [12] [13] suggested utilizing the cross attention mechanism to learn the inherent alignment between the encoded high-dimensional representations. Recently researchers turned to the large amount of unlabeled data, and adopt self-supervised audio-visual pretraining [14] [15] [16] [17] to narrow down the bias in data sources.

In this paper, we propose a new audio-visual speech recognition model with a unified cross-modal attention mechanism. The audio and visual input streams are processed at an early stage and mapped into a unified embedding space before encoding. Then the embeddings from different modalities are concatenated along the temporal dimension and can be modeled by a unified AVSR encoder. The self-attention module inside the encoders distinguishes different modalities with the help of respective positional encodings and additional modality embeddings. In this manner, the proposed model has at least two advantages. First, the backbone network can learn the cross-modal correlation based on the inter-modal and intra-modal context information with rare extra parameters dedicated to the auxiliary modality. Second, the unified self-attention for cross-modal modeling can naturally handle the asynchronized audio and video input that may appear in practical applications.

We conduct a series of experiments on the large-scale audio-visual dataset LRS3 [18]. The WER on clean utterances is further reduced compared to the state-of-the-art, and the performance under noisy conditions is improved significantly as well. Other evidence in the experiments also demonstrates the robustness and effectiveness of our model in special cases of modality corruption.

2. AUDIO-VISUAL SPEECH RECOGNITION

2.1. Baseline Models

The conformer encoder-decoder model has shown great performance on both the audio-only [7] and audio-visual [6] speech recognition. As shown in Fig.1.(a), the audio-visual model denoted as *Dual-Encoder* consists of separate front-ends and conformer encoders with a hybrid CTC/Attention architecture. In the acoustic front-end, audio input in waveform is transformed into filter-bank features and downsampled by a 2D-convolution block into features $x_a \in R^{t_a \times d}$. The visual front-end extract lip reading features [19] $x_v \in R^{t_v \times d}$ from the mouth ROIs in video clips by a 3D-convolution block, a ResNet34 network [20], and a two-layer Bi-LSTM. Here t_a and t_v are the number of frames in the acoustic and visual features but they are not necessarily equal, and d is the dimension of the feature

Yanmin Qian is the corresponding author.

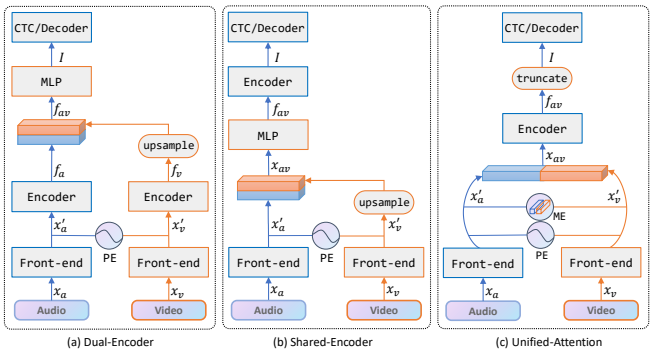

Fig. 1: Diagrams of different structures for audio-visual speech recognition. (Blue: traditional ASR workflow; Orange: auxiliary visual workflow.) (a) *Dual-Encoder* [6] encodes the audio-visual input separately and concatenates them along the channel dimension after forced alignment. (b) *Shared-Encoder* is a naive early fusion structure that fuses the audio-visual features before the single encoder. (c) *Unified-Attention* is the proposed structure with unified cross-modal attention. The audio-visual features are concatenated along the temporal dimension before the single encoder. *PE* and *ME* are short for Positional Encoding and Modality Embeddings.

space. Then they are encoded by the dedicated conformer encoders into representations $f_a \in R^{t_a \times d}$ and $f_v \in R^{t_v \times d}$ after adding the positional encoding. Representations of two modalities are concatenated along the channel dimension after the visual feature sequence gets upsampled to match the temporal length of the acoustic feature sequence. A multi-layer perceptron (MLP) is connected to project the doubled feature size back such that the output $I \in R^{t_a \times d}$ can be fed into the CTC and transformer decoder. This middle fusion strategy can be calculated as

$$f_a = \mathbf{Encoder}(\mathbf{PE}(x_a)), \tag{1}$$
$$f_v = \mathbf{Encoder}(\mathbf{PE}(x_v)), \tag{2}$$
$$f_{av} = f_a \oplus^c \mathbf{Upsample}(f_v), \tag{3}$$
$$I = \mathbf{MLP}(f_{av}), \tag{4}$$

where the $\oplus^c$ means concatenation along the channel dimension.

The audio-only model is a substructure of the audio-visual model, as the workflow marked in blue in Fig.1.(a), where all the blocks related to the visual modality are excluded.

2.2. Proposed Model

Instead of the middle fusion strategy in the baseline audio-visual model, we utilize the **unified cross-modal attention** mechanism to conduct an early fusion of the input audio-visual streams. As shown in Fig.1.(c), the proposed model denoted as *Unifed-Attention* comprises two front-ends and a single encoder for the fused audio-visual features. The shallow acoustic features $x_a \in R^{t_a \times d}$ and visual features $x_v \in R^{t_v \times d}$ after the front-ends are concatenated along the temporal dimension, forming a unified new sequence $x_{av} \in R^{(t_a+t_v)\times d}$ of both modalities. The positional encodings are calculated respectively for the subsequence of each modality and are concatenated in the same way to match the size of the unified sequence. Apart from the positional encoding, two learnable modality embeddings $\mathrm{ME}_a \in R^{1\times d}$ and $\mathrm{ME}_v \in R^{1\times d}$ are added to the acoustic and visual features correspondingly. With these two prior steps, the attention modules inside the encoder can distinguish the two modalities and stay aware of the relative order in each subsequence. Then the encoder operates directly on the unified sequence, which leads to a large and unconstrained context for cross-modal attention. At the end of the encoders, the visual frames are discarded while the subsequence of the audio modality is fed into the successive CTC and decoder since the cross-modality information is already well exchanged. In summary, the process can be formulated as

$$x'_a = \mathbf{PE}(x_a) + \mathbf{ME}_a, \tag{5}$$
$$x'_v = \mathbf{PE}(x_v) + \mathbf{ME}_v, \tag{6}$$
$$x_{av} = x'_a \oplus^t x'_v, \tag{7}$$
$$f_{av} = \mathbf{Encoder}(x_{av}), \tag{8}$$
$$I = f_{av}[1, \cdots, t_a], \tag{9}$$

where the $\oplus^t$ means concatenation along the temporal dimension. This fusion strategy requires few parameters for the visual stream and does not need explicit alignment between the audio-visual sequences.

The operations in the unified cross-modal attention mechanism are decoupled from the encoders, which makes it feasible for the model to generalize to the input sequence of only one single modality existing. Empirical results in the literature on lip reading [21] [22] and ASR [23] [7] have demonstrated that visual evidence is relatively weak for the speech recognition task, compared to the well-exploited acoustic signal. Therefore, we propose to optimize our audio-visual model in a mixed-type training style, which means the audio-only and audio-visual data can be both used for model training. Since all the data in LRS3 contains both audio and visual modalities, we randomly drop the visual data in the training iterations. With a probability of p_m, the modality dropout gate will make the input of visual modality neglected once in the forward and backward path, i.e., for the i-th iteration:

$$\mathbf{Input}_i = \begin{cases} \{\mathbf{Audio}_i, \mathbf{Video}_i\}, & \text{with } 1 - p_m \\ \{\mathbf{Audio}_i\}, & \text{with } p_m \end{cases} \tag{10}$$

In this manner, the training task will occasionally be downgraded to an ASR task that is less complex than the AVSR task. Thus the training style helps provide a solid optimization foundation for the audio-visual model, especially for the attention mechanism on the unified sequences.

3. EXPERIMENTS

3.1. Data Preparation

The model training and testing are conducted on the large-scale audio-visual dataset Lip Reading Sentences 3 (LRS3) [18]. The LRS3 dataset consists of 433 hours of utterances with their corresponding video clips and text transcriptions. The dataset is officially divided into *pretrain*, *trainval* and *test* sets. We separate 10,000 samples from the *trainval* set as the validation data, and the remaining samples are combined with those in the *pretrain* set as the training data. Any sample with a duration of less than 1 second is excluded from the dataset.

To simulate the real environment, we use noises from WHAM noise dataset [24] to generate noisy samples by mixing the clean utterances and noises with SNR uniformly sampled in $[-6, 6]$. Those noisy samples are used for training along with the clean ones. We also generate noisy samples using the validation/testing utterances

Table 1: WER(%) comparison of different models under various conditions. (†: extra training data.)

Model	Train	Test	SNR(dB) -10	-5	0	5	10	15	20	Noisy average	Clean
Conformer [7]	A	A	55.4	23.6	8.8	4.0	3.0	2.5	2.4	14.2	2.2
†*AVHubert*(*Base*) [15]	AV	AV	*21.1*	*9.6*	*4.7*	*2.9*	*2.3*	*2.2*	*2.1*	***6.4***	***2.0***
Shared-Encoder			30.3	14.2	7.5	5.1	4.6	4.3	4.1	10.0	4.0
Dual-Encoders [6]			32.6	14.4	7.1	4.1	3.1	2.6	2.5	9.5	2.3
Unified-Attention			24.6	10.5	5.3	3.3	2.6	2.3	2.3	**7.3**	**2.1**
w/o mixed-type training			25.1	10.9	5.2	3.7	2.7	2.6	2.5	7.5	2.4
†*AVHubert*(*Base*) [15]	AV	A	*79.3*	*42.6*	*16.0*	*7.4*	*4.8*	*3.5*	*2.9*	*22.4*	*2.6*
Dual-Encoders [6]			60.7	26.9	10.9	5.4	3.7	3.1	2.8	16.2	2.7
Unified-Attention			60.5	26.3	9.8	5.0	3.0	2.5	2.5	**15.7**	**2.4**

and the unseen noises with SNR in $\{-10, -5, 0, 5, 10, 15, 20\}$, and they are used for validation/testing along with the clean data.

3.2. Experimental Setups

Most of the hyper-parameters are shared among all the conformer-based models. In the acoustic front-end, the STFT has n-fft $= 512$, window-size $= 400$, and hop-length $= 160$. The encoder consists of 12 conformer blocks, with the hidden size of fully-connected layers $d^{fc} = 1024$, the hidden size in attention modules $d^{att} = 256$, the number of attention heads $n^{head} = 4$, and the kernel size in convolution modules $k = 31$. The decoder consists of 6 transformer blocks, with $d^{fc} = 2048$, $n^{head} = 6$, and the vocubulary size as 5000. We adopt relative positional encoding for the conformer blocks and absolute positional encoding for the transformer blocks. The p_m for mixed-type training is set empirically as 0.35.

The visual front-end is pretrained on the LRW dataset [25]. We extract the 512-dimension visual features in advance and exclude the visual front-end during training for acceleration. All the trainable parameters are initialized randomly and optimized by the Adam [26] optimizer with a warmup learning rate scheduler. The batch size is set dynamically as the number of elements in the acoustic input is fixed to 45 million. The models are trained with the peak learning rate as 0.002 at the 15,000-th step for 80 epochs. We use SpecAug [27] to augment the audio input. For the visual input, we adopt a similar strategy to augment the extracted feature vectors.

To obtain a strong language model, we use the texts from the LibriSpeech-960h dataset [28] along with texts from the LRS3 training data as the training corpus, which is consistent with the previous work [6]. A transformer-based language model [29] is trained for 25 epochs and we get a perplexity of 54 on the LRS3 testing data. The weight of the language model for decoding is empirically set as 0.2.

3.3. Results and Analysis

3.3.1. Comparison to the state-of-the-art

Apart from the audio-visual baseline model *Dual-Encoder* and the audio-only one mentioned in 2.1, we also test the data on other models to support the results. We implement a naive early fusion model denoted as *Shared-Encoder* as shown in Fig.1.(b). Audio-visual features are concatenated along the channel dimension ahead of the single encoder. This *Shared-Encoder* model has a similar number of parameters and also uses an early fusion strategy compared with our model. Moreover, we take for reference the publicly released AVHubert [15] model in *Base* version, which also has 12 encoding blocks but is pretrained with extra data from the VoxCeleb2 [30] dataset and then finetuned on the LRS3 dataset.

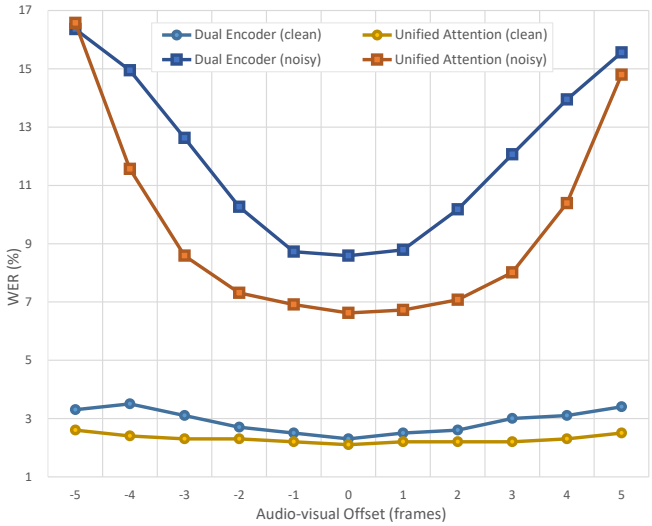

Fig. 2: WER variation comparison of the baseline and proposed models with different audio-visual offsets. A positive offset means the visual sequence is artificially shifted forward, and vice versa.

As shown in the central part of Table 1, our model reduces the Word Error Rate (WER) on clean samples from the state-of-the-art 2.3% to 2.1%, which is difficult to improve since lip movements are much more confusing than the clean utterances. If the mixed-type training is eliminated, the performance on noisy utterances will become slightly worse yet the WER on clean utterances will deteriorate seriously. The baseline *Dual-Encoder* model performs much better than the audio-only baseline model under noisy conditions because of the complementary effect of the visual modality, and yet our model improves the performance significantly again by relatively 23%, demonstrating the advantage of the proposed unified cross-modal attention mechanism on exploiting visual information. The naive *Shared-Encoder* model behaves badly for the clean data, but has a fair performance when the noise is extremely strong (SNR under 0dB), which indicates that the early fusion strategy is naturally suitable for audio-visual fusion. The performance of the pretrained AVHubert model is prominent as expected, but it's reason-

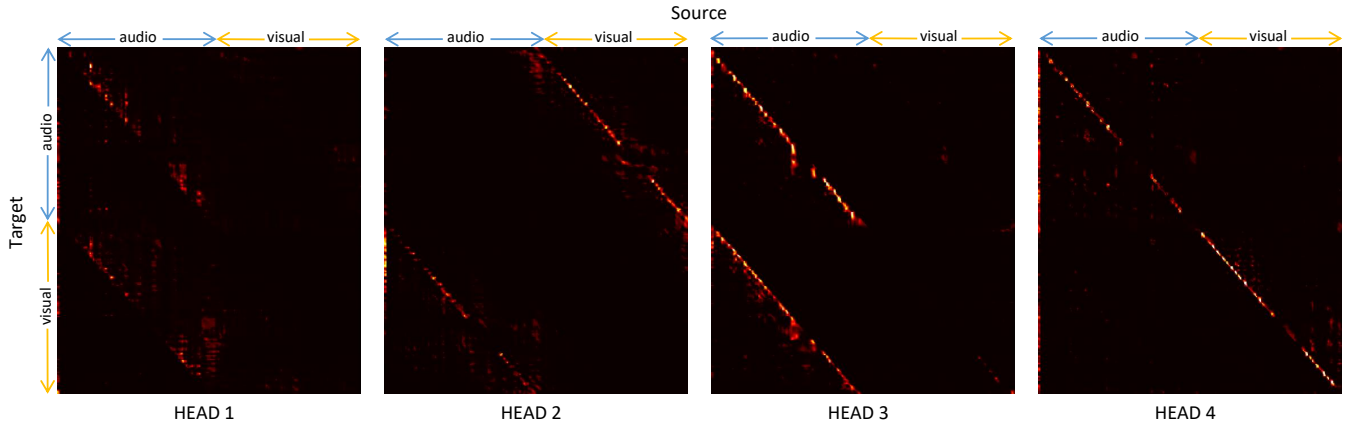

Fig. 3: Visualization of attention maps in the proposed unified cross-modal attention. They are from the four attention heads inside the last encoding block. Brighter pixels have larger attention weights. The input sample is randomly chosen from the testing data.

able due to the self-supervised training with a large amount of unlabeled data. This gap between the AVHubert model and our model is within our expectations, and we look forward to leveraging the unified cross-modal attention mechanism in self-supervised learning in future work.

3.3.2. *Inference in the absence of visual modality*

When the auxiliary visual modality is absent, which is common in real-world scenarios due to environmental disturbance or equipment fault, an audio-visual model is expected to keep working and bring a satisfying performance with the audio-only input. As illustrated in section2.2, the proposed model can be optimized with a mixed-type training style, and thus the model has the potential to inference the audio-only input data normally. Hence, we test the proposed model together with other models on the audio-only samples. Our model requires no modification to the input interface, but for other models, the visual modality should be filled with zeroes. We report the experimental results in the bottom part of Table 1. The proposed model gets the best scores compared with other models, and it is even much better than the pretrained AVHubert in this situation. It is prominent that on clean utterances, the WER can be still maintained at 2.4%, and the slight performance drop is also within the acceptable range compared to the fully audio-only baseline system.

3.3.3. *Roustness to audio-visual misalignment*

One of the natural advantages of the unified cross-modal attention mechanism is that there's no need to manually force a frame-level alignment on the audio-visual sequences. To verify this hypothesis, we measure the WER after artificially shifting the visual input sequence by a frame offset in $[-5, 5]$, i.e. from about -200 to 200 ms. The chart in Fig.2 shows the performance variation of our proposed new model and the *Dual-Encoder* baseline model. For the bottom part from clean testing samples, our model keeps a stable performance even though there's no such augmentation of misalignment in the training data, while the performance of the *Dual-Encoder* model jitters with the offset changing. For the top part from the noisy testing samples, both models drop to the same level of WER when the offset is large enough, which means severe misalignment will totally corrupt the validity of the visual modality. However, for the offsets in $[-3, 3]$, our model doesn't degrade as seriously as the baseline model.

3.3.4. *Visualization of the unified cross-modal attention*

To give a qualitative assessment of the unified cross-modal attention mechanism, we visualize the attention maps from the multi-head attention modules in Fig.3. The four attention heads come from the last encoding block, and they behave quite differently. Apart from the normal attention inside each modality, there exists cross-modal attention to exchange information. And it's worth mentioning that, the acoustic features are shown to be more emphasized in the attention module as the audio regions are more frequently attended to in the attention maps. This is consistent with the intention of the mixed-type training style to keep the audio dominant and the video auxiliary. The inter-modality and intra-modality interactions indicate that the model is able to learn an implicit audio-visual alignment on the contextual level.

4. CONCLUSION

In this paper, we have explored a new fusion mechanism for the audio-visual speech recognition (AVSR) task, where the input sequences from both modalities are concatenated along the temporal dimension in the unified space at an early stage of the model. The large and flexible cross-modal context with the mixed-type training style facilitates the adaptive fusion of the audio-visual information. On the large-scale LRS3 dataset, the proposed model with unified cross-modal attention significantly improves the performance in both clean and noisy environments compared to the state-of-the-art. Experiments also demonstrate our model's robustness to the possible absence of the visual modality or frame misalignment in audio-visual streams. We expect to incorporate this mechanism with other advanced techniques to further boost the capability of AVSR.

5. ACKNOWLEDGMENTS

This work was supported in part by China NSFC projects under Grants 62122050 and 62071288, and in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102.

6. REFERENCES

[1] Yajie Miao and Florian Metze, "Open-domain audio-visual speech recognition: A deep learning approach," in *Proc. ISCA Interspeech*, 2016, pp. 3414–3418.

[2] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G. Okuno, and Tetsuya Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[3] Satoshi Tamura, Hiroshi Ninomiya, Norihide Kitaoka, Shin Osuga, Yurie Iribe, Kazuya Takeda, and Satoru Hayamizu, "Audio-visual speech recognition using deep bottleneck features and high-performance lipreading," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. 2015, pp. 575–582, IEEE.

[4] Triantafyllos Afouras, Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman, "Deep audio-visual speech recognition," *CoRR*, vol. abs/1809.02108, 2018.

[5] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic, "Audio-visual speech recognition with a hybrid ctc/attention architecture," in *Proc. IEEE SLT*. 2018, pp. 513–520, IEEE.

[6] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "End-to-end audio-visual speech recognition with conformers," in *Proc. IEEE ICASSP*. 2021, pp. 7613–7617, IEEE.

[7] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. ISCA Interspeech*, 2020, pp. 5036–5040.

[8] Jianwei Yu, Shi-Xiong Zhang, Jian Wu, Shahram Ghorbani, Bo Wu, Shiyin Kang, Shansong Liu, Xunying Liu, Helen Meng, and Dong Yu, "Audio-visual recognition of overlapped speech for the LRS2 dataset," in *Proc. IEEE ICASSP*. 2020, pp. 6984–6988, IEEE.

[9] Pan Zhou, Wenwen Yang, Wei Chen, Yanfeng Wang, and Jia Jia, "Modality attention for end-to-end audio-visual speech recognition," in *Proc. IEEE ICASSP*. 2019, pp. 6565–6569, IEEE.

[10] George Sterpu, Christian Saam, and Naomi Harte, "How to teach dnns to pay attention to the visual modality in speech recognition," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1052–1064, 2020.

[11] George Sterpu, Christian Saam, and Naomi Harte, "Attention-based audio-visual fusion for robust automatic speech recognition," in *Proc. the 20th ACM International Conference on Multimodal Interaction*. 2018, pp. 111–115, ACM.

[12] Liangfa Wei, Jie Zhang, Junfeng Hou, and Lirong Dai, "Attentive fusion enhanced audio-visual encoding for transformer based robust speech recognition," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. 2020, pp. 638–643, IEEE.

[13] Yifei Wu, Chenda Li, Song Yang, Zhongqin Wu, and Yanmin Qian, "Audio-visual multi-talker speech recognition in a cocktail party," in *Proc. ISCA Interspeech*, 2021, pp. 3021–3025.

[14] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *International Conference on Learning Representations*, 2022.

[15] Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed, "Robust self-supervised audio-visual speech recognition," *CoRR*, vol. abs/2201.01763, 2022.

[16] Zi-qiang Zhang, Jie Zhang, Jian-Shu Zhang, Ming-Hui Wu, Xin Fang, and Li-Rong Dai, "Learning contextually fused audio-visual representations for audio-visual speech recognition," *CoRR*, vol. abs/2202.07428, 2022.

[17] Xichen Pan, Peiyu Chen, Yichen Gong, Helong Zhou, Xinbing Wang, and Zhouhan Lin, "Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition," in *Proc. ACL*. 2022, pp. 4491–4503, Association for Computational Linguistics.

[18] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," *CoRR*, vol. abs/1809.00496, 2018.

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

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*. 2016, pp. 770–778, IEEE Computer Society.

[21] Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman, "Lip reading sentences in the wild," in *Proc. IEEE CVPR*. 2017, pp. 3444–3453, IEEE Computer Society.

[22] Brendan Shillingford, Yannis M. Assael, Matthew W. Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, Marie Mulville, Misha Denil, Ben Coppin, Ben Laurie, Andrew W. Senior, and Nando de Freitas, "Large-scale visual speech recognition," in *Proc. ISCA Interspeech*, 2019, pp. 4135–4139.

[23] Titouan Parcollet, Mohamed Morchid, and Georges Linarès, "E2E-SINCNET: toward fully end-to-end speech recognition," in *Proc. IEEE ICASSP*. 2020, pp. 7714–7718, IEEE.

[24] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux, "Wham!: Extending speech separation to noisy environments," in *Proc. ISCA Interspeech*, 2019, pp. 1368–1372.

[25] Joon Son Chung and Andrew Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision (ACCV)*. Springer, 2016, pp. 87–103.

[26] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[27] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. ISCA Interspeech*, 2019, pp. 2613–2617.

[28] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE ICASSP*. 2015, pp. 5206–5210, IEEE.

[29] Kazuki Irie, Albert Zeyer, Ralf Schlüter, and Hermann Ney, "Language modeling with deep transformers," in *Proc. ISCA Interspeech*, 2019, pp. 3905–3909.

[30] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. ISCA Interspeech*, 2018.