INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

An Improved End-to-End Audio-Visual Speech Recognition Model

Sheng Yang, Zheng Gong, Jia Kang*

College of Computer Science, Inner Mongolia University, National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, Inner Mongolia Key Laboratory of Mongolian Information Processing Technology, Hohhot, China

ysbeard@163.com, csgzh@imu.edu.cn, 1294722198@qq.com

Abstract

By incorporating lip language, audio-visual speech recognition can effectively improve the recognition effect in noisy environments, and will slightly improve the recognition effect in quiet environments. we use a frequency domain attention based residual network (Fca-Net) as the model of the vision front-end module, which extracts more features that are helpful to the AVSR and VSR system at a small cost. And use the powerful speech pre-training model Hu-BERT as the recognition front-end model of ASR. We compare the impact of different model as visual back-end modules and fusion modules on the AVSR system. Our experiments show that the model selection of the fusion module is critical to the performance of the AVSR system. Ultimately, our proposed model achieves state-of-the-art results on audio-visual speech recognition tasks using the LRS2 dataset.

Index Terms: audio-visual speech recognition, frequency domain attention, Hu-BERT, feature fusion

1. Introduction

Audio-visual speech recognition (AVSR) is a multi-modal speech recognition task that combines audio and visual streams. Since visual streams mainly rely on lip features to identify corresponding text, there will be problems with the same mouth shape but different pronunciations, resulting in the accuracy of lip language recognition lower. The audio stream is also affected by noise. It is still challenging to improve the accuracy of the two, and to effectively combine the two to achieve complementary effects.

Traditional speech recognition uses manual features and machine learning for recognition. *Dupont et al.* [1] added handcrafted visual features on this basis to improve the accuracy of speech recognition. With the development of learning, manually crafted features are gradually replaced by features extracted through deep learning. In ASR, *Parcollet et al.* [2] and N. *Zeghidour et al.* [3] have demonstrated that features extracted through deep learning achieve better results. With the popularity of transformers [4], *Gulati et al.* [5] designed a convolution-enhanced transform (Conformer) and achieved good results. In VSR, *Stafylakis et al.* [6] based on *Assael et al.* [7] have achieved good results by using a deeper network and proposing a 3D-ResNet network structure. At the same time, *Martinez et al.* [8] designed a network structure based on time convolution (TCN) for prediction, and *Ma et al.* [9, 10] designed a more efficient TS-TCN for prediction and improved it using different training strategies. *Kim et al.* [11] added a multi-head attention mechanism to improve lip reading for homonyms Word recognition effect. *Prajwal et al.* [12] designed the coding layer of the Visual Transformer Pooling (VTP) structure, which has led to a new level of accuracy in lip language recognition.

Some recent AVSR systems using deep learning have achieved promising results. For example, *Afouras et al.* [13] proposed a model using Seq2Seq as a loss function, while *Petridis et al.* [14] designed a hybrid CTC/attention architecture. *Ma et al.* [15] proposed an end-to-end audiovisual speech recognition model based on Conformer. In recent years, self-supervised pre-training models have attracted attention. *Shi et al.* [16] designed a self-supervised pre-training model AV-Hubert, while *Pan et al.* [17] used self-supervised models as single-mode state front-end input and achieved good results on LRS2[18]. The audio front-end uses a self-supervised learning pre-training model called Wav2vec [19], and the visual front-end uses MOCO [20,21], a self-supervised model of contrastive learning. However, the generalization of MOCO to lip images may not be as good since it is pre-trained on ImageNet [22]. In our model, an advanced speech pre-training model will be used, and the image pre-training model will not be used, Specifically, our contributions are the following:

1.In the visual front-end, we use the Res-Net [23] network based on the frequency-domain attention mechanism developed by *Qin et al.* [24]. This model extracts important visual features from the frequency domain, and improves the model's effectiveness at a small cost in terms of added parameters. In addition, we use the advanced Hu-BERT [25] as the audio pre-training model for the audio front-end.

2.We explore the impact of using Transformer and Conformer as the visual back-end models on VSR and AVSR systems.

3.We compare the effectiveness of using MLP, Transformer encoder, and Conformer as the models for the fusion part. Our experiments show that the transformer provides the good results for feature fusion.

2. Methodology

2.1. Visual Front-end

For the visual flow, we use Fca-Net-50[24] to extract the features of the 112*112 mouth grayscale image. This model converts the feature map of each channel of ResNet-50 into a frequency domain map and selects the value of a certain position, puts the value into the MLP layer for learning, and finally obtains the attention weight of the channel. Simply put, the model weights the output of each layer of ResNet from the perspective of the frequency domain. However, there is not much difference between the front and rear video frames of the lips. Simply using one frame as the recognition target will lose the features brought about by the changes of the lips. Therefore, 3D convolution is added before Fca-Net[24] to capture lip shape features and continuous changing characteristics.

10.21437/Interspeech.2023-2256

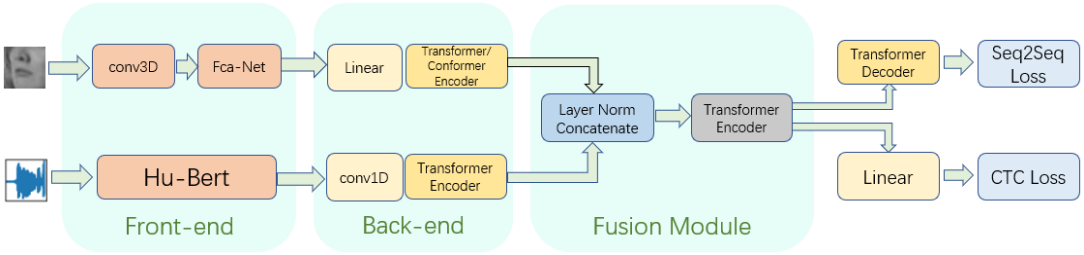

Figure 1: *This picture shows the overall structure diagram of the model, which is divided into 4 parts, which are the front-end modules of visual and auditory, back -end modules, fusion modules, decoding modules*

2.2. Audio Front-end

For the auditory stream, we use Hubert, which is advanced than Wav2vec, as the model of the Audio Front-end. In Hubert, firstly a clustering model is obtained through unsupervised training to discretize the speech signal and obtain the target sequence (Hidden units). Then, the MLM self-supervised pre-training method similar to BERT is used to enable the model to predict the target value of the mask position through the speech signal after masking.

2.3. Visual back-end

The conformer encoder and transformer encoder are respectively used as the visual back-end model. However, the standard conformer encoder model has a downsampling process that reduces the sequence length of the feature to 25% due to the video frame rate of 25FPS. If the frequency is reduced too much, the number of features within a given time interval may be less than the predicted number of tokens, which can make decoding challenging. To address this, we remove the downsampling process of the conformer encoder to ensure that the frequency of the features is greater than the predicted number of tokens.

2.4. Audio back-end

As the audio front-end pre-training model is already powerful, there is no need to make extensive comparisons with the audio back-end models, and only the Transformer encoder is used for encoding. Since the output of Hubert is 49Hz, which is twice the sampling rate of the visual modality, a 1-D convolution layer with a stride of 2 is added to the Transformer encoder to ensure that the frequency of the output features of both modalities is the same in the back-end output.

2.5. Fusion

Front-end and Audio back-end output are both 512-dimensional feature vectors with a frequency of 25. After batch normalization and merging, 1024-dimensional feature vectors can be obtained. For the new feature vectors, three types are used. The deep fusion of the neural network is the stacking of multiple MLP, the Conformer encoder and the Transformer encoder.

2.6. Decoder

The first type of Decoder uses the transformer seq2seq decoder for decoding, including the basic block of the 6-layer transformer. During training, teacher forcing is performed at the character level using ground truth characters as input, and are trained with cross-entropy loss. The second type relies on the linear layer of CTC loss for training and decoding, including 4 linear layers and the corresponding ReLU activation function. The output is the CTC posterior probability of each input frame. Experiments by *Afouras et al.* [13] show that Seq2Seq is better than CTC, but the difference in performance is not significant. We use the two in combination, Following the setting of *Petridis et al.* [14] two decoders trained simultaneously based on the same output in the fusion module.

2.7. Loss fusion

Reference *Petridis et al.* [14] use a hybrid CTC/attention loss during training. Suppose $x = [x_1, \cdots, x_T]$ is the output sequence of the fusion model, and $y = [y_1, \cdots, y_L]$ is the targets corresponding to the frame, with T and L representing the input and target lengths.

Firstly, the CTC loss assumes that each output prediction is conditionally independent and has the following form, where x is the output of linear and y is the true label corresponding to the feature.

$$p_{CTC}(y|x) \approx \prod_{l=1}^{L} p(y_l|x) \tag{1}$$

Secondly, an attention-based model removes this assumption by directly estimating the posterior probability based on the chain rule, which can be expressed as the following formula, where x is the output of Transformer Encoder and y is the true label corresponding to the feature.

$$p_{CE}(y|x) \approx \prod_{l=1}^{L} p(y_l|y < l, x) \tag{2}$$

The final loss calculation formula combines the two loss functions, with a weight coefficient α for CTC and attention mechanisms. This hybrid CTC/attention loss function has also been applied in automatic speech recognition (ASR) and visual speech recognition (VSR).

$$\mathcal{L} = \alpha log p_{CTC}(y|x) + (1-\alpha) log p_{CE}(y|x) \tag{3}$$

3. Experiment

3.1. Datasets

In this work, our training and validation are mainly conducted on the LRS2 dataset [18]. We also utilize the LRW dataset [26] for Curriculum Learning to train the visual front-end. This approach significantly improves the convergence speed of the loss during the training of the visual-only (VO) model.

3.2. Data Pre-processing and augmentation

In each video, dlib is used to detect and track 68 facial landmarks. Subsequently, a bounding box of size 120×120 is used to crop the mouth regions of interest (ROIs). then converted to grayscale and normalized with respect to the overall mean and variance on the training set. To address the issue of overfitting and enhance the model's ability to generalize, we adopted a

technique proposed by *Pan et al.* [17] where each preprocessed image sequence is randomly cropped to a size of 112×112 and horizontally flipped with a probability of 0.5 during pre-training and training. In our experiments, we observed that using a smaller mouth region during testing can slightly improve the accuracy of lip recognition. Therefore, during testing, we crop a region with dimensions of 88×88 from the center of the mouth.

Each raw audio waveforms are normalized to zero mean and unit variance by subtracting its mean and dividing by its standard deviation. During audio-only training, additive noise is introduced in the time-frequency domain of the original audio waveform. Babble noise is added to the original audio stream with an SNR level ranging is 5dB and a probability of $p_n = 0.25$. Babble noise is synthesized by mixing 20 different audio samples from the LRS2 dataset.

3.3. Experimental hyperparameter settings

In this work, the parameters of the Fca-Net-34 model trained on ImageNet are used as the initial parameters for pre-training on the LRW dataset. The pre-trained model's parameters are then used for the front-visual-end model, while the rest of the model parameters are randomly initialized. Each stage of model training used the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The learning rate is initialized to 10^{-4}, which warm up and reduce on plateau scheduler and a final learning rate of 10^{-5}. And all the models we used dropout with p =0.1 and label smoothing. The relative weight in CTC loss and seq2seq loss λ is set to 0.2. The Fca-Net model utilizes 16 pre-selected frequency domain positions, as described in the paper by *Qin et al.* [24] All Transformer encoders use a set of hyperparameters (num_layer = 6, heads_num = 8, dff = 2048, d_model = 512), All Conformer encoders use the same set of hyperparameters from the original paper's medium Conformer (num_layer = 16, heads_num = 4, encoder_dim = 256, d_model = 512, conv_kerne-l_size = 32). Our implementation is based on the Pytorch library and trained on three NVIDIA V100 GPUs with a total of 96GB memory for 2 weeks.

3.4. Evaluation

For all of our experiments, we measure the performance using the Word Error Rate (WER), which is defined as WER = (S + D + I) / N. Here, S, D, and I represent the number of substitutions, deletions, and insertions needed to transform the hypothesis into the reference transcript, and N is the total number of words in the reference transcript.

4. Results

4.1. Result in LRS2

Table 1 presents the results of our model and other models in the Visual-only, Audio-only, and Audio-Visual. Notably, our model does not utilize any language model, whereas other models leverage language models trained by NN, RNN, and Transformer. Moreover, our model solely utilizes the LRW dataset as external data, whereas other models rely on LRS3[27], LSVSR[28], and other datasets as external data.

4.1.1. Only-visual-Result

The data used for training includes labeled LRS2 data in its pre-train and train sets and the LRW data used in Curriculum learning. The visual-only model achieved a WER of 37.7%, lagging behind the current state-of-the-art VTP with 15.1%. The VTP model uses more training datasets, including a large non-public dataset MV-LRS[32], while we only use the LRW dataset and the LRS2 dataset. Several experiments have proved that using more datasets will significantly improve the accuracy of the speech recognition model, and the front-end model of VTP uses a more complex structure that greatly increases the number of parameters and the training time, which has reached 14 days. In contrast, the front-end of our model uses Fca-Net, which only adds a channel attention mechanism in the middle of each block, equivalent to adding an MLP layer with a small number of parameters in the middle of each block. Furthermore, the token predicted by VTP is WordPiece, while our token is 40 characters, which is more difficult. Our model is more suitable for comparison with the end-to-end Conform model, because it has the same number of parameters as our model, and the predicted token is the same as the data set used. The error rate was reduced by 0.2%. Compared with the end-to-end Conform without language model, the error rate is reduced by 4.7%.

Table 1: *Audio-only, visual-only and audio-visual results of word error rate (WER) tested on LRS2. Models with an * denote that results are using an external language model.*

Methods	WER
Visual-only	
TM-CTC*[13]	54.7
Conv-seq2seq[29]	51.7
TM-seq2seq*[13]	50.0
KD-TM[30]	49.2
LF-MMI TDNN*[31]	48.9
E2E Conformer*[15]	37.9
MOCO+Wav2vec [17]	43.2
VTP[12]	**22.6**
Our Model(transformer)	38.3
Our Model(conformer)	37.7
Audio-only	
TM-CTC*[13]	10.1
TM-seq2seq*[13]	9.7
CTC/attention*[14]	8.2
LF-MMI TDNN*[31]	6.7
E2E Conformer*[15]	3.9
MOCO+Wav2vec [17]	2.7
Our Model	**2.2**
Audio-Visual	
TM-seq2seq*[13]	8.5
TM-CTC*[13]	8.2
LF-MMI TDNN*[31]	5.9
E2E Conformer*[15]	3.7
MOCO+Wav2vec [17]	2.6
Our Model	**2.1**

4.1.2. Only-audio-Result

In the primary audio setting, the pre-train and train sets in LRS2 are used as the train set in the Only-audio training stage, as well as the 60K hours unlabeled data from LibriLight [33] that are indirectly used through inheriting HuBERT parameters. Our model achieves a WER of 2.2%, which is a 0.5% reduction in WER compared to the current state-of-the-art [17], indicating a relative improvement of 19%.

4.1.3. Audio-visual-Result

The training data used for training the audio-visual model consists of 224 hours of labelled video data from the pre-train and

train sets in LRS2. Our proposed audio-visual model achieves a WER of 2.1% without the help of an external language model, which represents an improvement of 0.5% over the current state-of-the-art[17] and a relative improvement of around 19%.

4.2. Ablation Studies

4.2.1. Fca-net Contribution in Visual Word Classification:

Results of visual word classification on LRW are shown in Table 2. We train a model by replacing the front-end model of ResNet-50 initialized with MoCo v2 weights with Fca-Net-50. An additional absolute improvement of 1.8% was observed, demonstrating that the frequency-domain attention mechanism can enhance the model's performance.

Table 2: *Ablation study on visual word classification performance on LRW.*

Methods	WER
Baseline	74.6%
+ ResNet-50 front-end	76.7%
+ MoCo v2 front-end	79.0%
+Fca-Net-50	80.8%

4.2.2. Performance Breakdown in Visual-only Set-ting

Results of the visual-only model on LRS2 are shown in Table 3. Starting from *Afouras et al.*[13]we first introduce end-to-end training by using a hybrid CTC/attention decoder, resulting in an absolute improvement of 14.5%. Then adding the Curriculum learning on the LRW dataset, the result is a relatively improved by 2.8%, Following the design of *Ma et al.* [15], replacing the back end with Conformer, the result was further improved by 3.8%. Finally, adding the visual front-end to the ResNet based on the frequency-domain attention mechanism (Fca-Net), the WER was 37.7%. This leads to an absolute improvement of 4.7 %.

Table 3: *Ablation study on visual-only model on LRS2.*

Methods	WER
Baseline	63.5%
+ Hybrid CTC/attention	49.0%
+LRW pre-training	46.2%
+Conformer encoder	42.4%
+Fca-Net	37.7%

4.2.3. Performance Breakdown in Fusion

The results of the audio-visual model on LRS2 are shown in Table 4. In order to investigate the impact of different models on the fusion results, we conducted numerous experiments for comparison. It can be seen that regardless of whether the visual model backend is a Transformer or a Conformer, the fusion model using multiple MLP-stacked FC layers and Conformer encoders did not perform very well. In fact, the results were even worse than those of the pure audio model. However, when both fusion model and vision backend model use Transformer encoder, the performance of the audio-visual speech recognition model was 0.1% better than that of the single audio model. When adding 0dB noise, only using Transformer as the fusion model can lead to much better results than audio-only models. Analyzing the reason, the simple structure of MLP is not enough to effectively combine the features of the two modalities, it may use visual features as interference features. And the convolutional structure inside the Conformer may not be suitable for feature fusion, although it can or does a good job in single-modal speech recognition.

Table 4: *The results obtained by the experiment in different visual back-end models using different fusion models*

Visual back-end Model	Fusion Model	WER 0dB	WER clean
Conformer	FC	25.9%	2.4%
	Transformer	23.5%	2.2%
	Conformer	24.7%	2.3%
Transformer	FC	23.5%	2.4%
	Transformer	**19.5%**	**2.1%**
	Conformer	22.7%	2.3%

4.2.4. Robustness under Noisy Inputs

Results of the audio-visual model on LRS2 are shown in Table 4. In order to explore the speech recognition effect of audio-only and audio-visual models in the presence of noise, we add babble noise with different SNR to the audio, and the signal-to-noise ratios are 0dB and 5dB respectively. When the SNR level is 0dB, our audio-only and audio-visual models achieve WER of 22.3% and 19.5%, respectively, which is 9.8% and 5% higher than the results of MoCo+Wav2vec [17] When the SNR level is 5dB, our audio-only and audio-visual models achieve WER of 5.7% and 5.7%, respectively, which is 1.2% and 0.6% higher than the results of MoCo+Wav2vec [17].

Table 5: *Word error rate (WER) under different SNR levels. The noises are synthesized babble noises.*

Model	Modal	0 dB	5 dB	clean
TM-CTC[13]	AO	58.0%	-	10.5%
	AV	33.5%	-	9.4%
MoCo+Wav2vec [17]	AO	32.5%	6.8%	2.7%
	AV	24.5%	6.3%	2.6%
Our Model	AO	22.3%	**5.7%**	2.2%
	AV	**19.5%**	5.8%	**2.1%**

5. Discussion

We design a new audio-visual speech recognition model and achieve a word error rate of 2.1% on the audio-visual speech recognition task on LRS2 data. According to Table 1, it can be seen that as the WER of audio-only continues to decrease, the importance of lip features in the audio-visual speech recognition model is also getting lower and lower. We improve the performance of visual-only and use a variety of fusion models , are trying to increase the role of lip features in the AVSR system, thereby improving the accuracy of audio-visual speech recognition. In addition, due to the large size of the LRS2 dataset and the high training cost, all positions in the frequency domain were not experimented with in sequence to select the suitable frequency domain position for lips.

6. Ackonwledeg

This research is supported by the Inner Mongolia Natural Science Foundation (No.2020MS06018), Applied Technology Research and Development Program of Inner Mongolia Autonomous Region (No. 2021GG0158)

7. References

[1] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *in IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141-151, Sept. 2000.

[2] T. Parcollet, M. Morchid and G. Linarès, "E2E-SINCNET: Toward Fully End-To-End Speech Recognition," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 7714-7718.

[3] N. Zeghidour, N. Usunier, et al., "End-to-end speech recognition from the rawwaveform," *in Interspeech*, 2018, pp.781–785.

[4] A. Vaswani, N. Shazeer, et al., "Attention is all you need," *in NeurIPS*, 2017, pp. 5998–6008.

[5] A. Gulati, J. Qin, et al., "Conformer: Convolution-augmented transformer for speech recognition," *in Interspeech*, 2020, pp. 5036– 5040.

[6] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," *in Interspeech*, 2017, pp. 3652–3656.

[7] Y. M. Assael, B. Shillingford, et al., "Lipnet: End-to-end sentence-level lipreading," arXiv preprint arXiv:1611.01599, 2016.

[8] B. Martinez, P. Ma, et al., "Lipreading Using Temporal Convolutional Networks," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 6319-6323.

[9] P. Ma, B. Martinez, et al., "Towards Practical Lipreading with Distilled and Efficient Models," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, 2021, pp. 7608-7612.

[10] P. Ma, Y. Wang, et al., "Training Strategies for Improved Lip-Reading," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022, pp. 8472-8476.

[11] M. Kim, J. H. Yeo, and Y. M. Ro, "Distinguishing Homophenes Using Multi-Head Visual-Audio Memory for Lip Reading", *in AAAI*, vol. 36, no. 1, 2022, pp. 1174-1182.

[12] K. R. Prajwal, T. Afouras and A. Zisserman, "Sub-word Level Lip Reading With Visual Attention," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 5152-5162.

[13] T. Afouras, J. S. Chung, et al., "Deep Audio-Visual Speech Recognition," *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8717-8727, 1 Dec. 2022.

[14] S. Petridis, T. Stafylakis, et al., "Audio-visual speech recognition with a hybrid CTC/attention architecture," *2018 IEEE Spoken Language Technology Workshop (SLT)*, Athens, Greece, 2018, pp. 513-520.

[15] P. Ma, S. Petridis and M. Pantic, "End-To-End Audio-Visual Speech Recognition with Conformers," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, 2021, pp. 7613-7617.

[16] B. Shi, W.-N. Hsu, et al., "Learning audio-visual speech representation by masked multimodal cluster prediction", *Proc. Int. Conf. Learn. Representations*, 2022.

[17] X. Pan, P. Chen, et al., "Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition," *in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 4491–4503.

[18] J. S. Chung, A. Senior, et al., "Lip Reading Sentences in the Wild," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 3444-3453.

[19] Schneider, Steffen, et al., "wav2vec: Unsupervised Pre-Training for Speech Recognition, " *in Interspeech, 2019*, pp.3465-3469.

[20] X. Chen, H. Fan, et al. "Improved baselines with momentum contrastive learning." arXiv preprint arXiv:2003.04297, 2020.

[21] K. He, H. Fan, et al., "Momentum contrast for unsupervised visual representation learning", i*n 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729-9738, 2020.

[22] J. Deng, W. Dong, et al., "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 248-255.

[23] K. He, X. Zhang, et al., "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778.

[24] Z. Qin, P. Zhang, et al., "FcaNet: Frequency Channel Attention Networks," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 2021, pp. 783-792.

[25] W. -N. Hsu, Y. -H. H. Tsai, et al., "Hubert: How Much Can a Bad Teacher Benefit ASR Pre-Training?" *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, 2021, pp. 6533-6537.

[26] J. S. Chung and A. Zisserman, "Lip reading in the wild", P*roc. Asian Conf. Comput. Vis.*, pp. 87-103, 2016.

[27] T. Afouras, J. S. Chung, et al., "LRS3-TED: A large-scale dataset for visual speech recognition," in arXiv preprint arXiv:1809.00496, 2018.

[28] B. Shillingford, Y. Assael, et al., "Large-Scale Visual Speech Recognition," *in Proc. Interspeech 2019*, 2019, pp. 4135–4139.

[29] X. Zhang, F. Cheng and W. Shilin, "Spatio-Temporal Fusion Based Convolutional Sequence Learning for Lip Reading," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 2019, pp. 713-722.

[30] S. Ren, Y. Du, et al., "Learning from the Master: Distilling Cross-modal Advanced Knowledge for Lip Reading," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2021, pp. 13320-13328.

[31] J. Yu et al., "Audio-Visual Recognition of Overlapped Speech for the LRS2 Dataset," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 6984-6988.

[32] T. Afouras, J. S. Chung, et al., "Deep Audio-Visual Speech Recognition," *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, 2018, pp. 8717-8727.

[33] J. Kahn, M. Riviere, et al., "Libri-Light: A Benchmark for ASR with Limited or No Supervision," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 7669-7673.