Interspeech 2024
1-5 September 2024, Kos, Greece

SyncVSR: Data-Efficient Visual Speech Recognition with End-to-End Crossmodal Audio Token Synchronization

Young Jin Ahn[1⋆], *Jungwoo Park*[2⋆], *Sangha Park*[3], *Jonghyun Choi*[4], *Kee-Eung Kim*[1]

[1]KAIST [2]Kwangwoon University [3]Ajou University [4]Seoul National University

snoop2head@kaist.ac.kr, affjljoo3581@kw.ac.kr, wkrtkd222@ajou.ac.kr, jonghyunchoi@snu.ac.kr, kekim@kaist.ac.kr

Abstract

Visual Speech Recognition (VSR) stands at the intersection of computer vision and speech recognition, aiming to interpret spoken content from visual cues. A prominent challenge in VSR is the presence of homophenes—visually similar lip gestures that represent different phonemes. Prior approaches have sought to distinguish fine-grained visemes by aligning visual and auditory semantics, but often fell short of full synchronization. To address this, we present SyncVSR, an end-to-end learning framework that leverages quantized audio for frame-level crossmodal supervision. By integrating a projection layer that synchronizes visual representation with acoustic data, our encoder learns to generate discrete audio tokens from a video sequence in a non-autoregressive manner. SyncVSR shows versatility across tasks, languages, and modalities at the cost of a forward pass. Our empirical evaluations show that it not only achieves state-of-the-art results but also reduces data usage by up to ninefold.

Index Terms: visual speech recognition, lip-reading, crossmodal learning, end-to-end training, data efficiency

1. Introduction

Visual Speech Recognition (VSR), also referred to as lip-reading, constitutes the process of decoding spoken language through the observation of the visual cues, specifically the movements of the lips and facial dynamics. This technology holds critical importance in a variety of contexts, including the interpretation of lip movements from individuals with speech disorders [1], benefiting individuals with hearing disorders [2], recognizing spoken content within environments where acoustic signals are compromised [3,4], providing voiceovers to silent historical films, and fortifying security systems [5].

The primary challenge encountered in VSR stems from the inherent scarcity of information that can be extracted from visual cues alone [7]. Central to this issue is the presence of homophenes, wherein disparate sounds are visually manifested through identical or nearly identical lip movements [8]. Such phenomena represent significant ambiguity in the analysis of visemes, the fundamental units of visual speech recognition. This ambiguity poses considerable difficulties, as it muddles the clarity of speech interpretation through visual means alone.

Previous research to overcome such limitations has predominantly revolved around aligning visual with auditory semantics, attempting to reduce the gap between audio models and visual models. Earlier techniques [9–11] harnessed knowledge distillation from pretrained Automatic Speech Recognition (ASR) systems. Subsequent studies [8, 12, 13] trained auxiliary audio modules along with the visual encoder in order to transfer speech knowledge. However, the aforementioned methods create indirect links to the acoustic data, as visual models interact with the semantics of audio encoders rather than the acoustic data. Such audio modules might provide insufficient hidden knowledge to the visual modules due to the crossmodal gap in representations [7]. Recent works [14–16] strived to connect the visual encoder with speech data, but they utilized handcrafted features (e.g., spectrograms or MFCCs) as their inputs or targets, which possibly encompass inductive biases that could affect the learned representations [17].

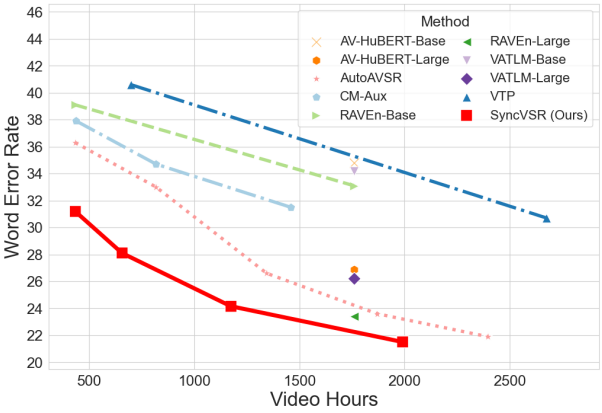


Figure 1: *Performance of SyncVSR on LRS3 [6] benchmark. SyncVSR outperforms available methods given the similar amount of video data resources. Our method also advances a tier in model size, where our base-size model shows superior performance compared to other large-size models.*

Moreover, several works [15–18] introduced learning methods based on crossmodal masked reconstruction, where portions of visual inputs are replaced with masked frames and models are trained to reconstruct corresponding audio representations. Nevertheless, an alternative method to masked segment reconstruction has been introduced in the Natural Language Processing (NLP) domain, which is token-level discrimination [19, 20]. A key advantage of such discriminatory supervision is that the model learns from all input tokens instead of just the small masked-out subset, advancing a tier of performance and being more sample-efficient [20]. This method is even more promising in the VSR domain since there is a fine-grained correspondence between the visual and auditory modalities, which provides a natural source of self-supervision [17].

In light of the above, we propose the SyncVSR framework, an innovative approach to VSR that directly aligns visual phonetic units with their acoustic counterparts through quantized audio tokens, facilitating robust end-to-end crossmodal syn-

⋆ Equal contribution.

10.21437/Interspeech.2024-432

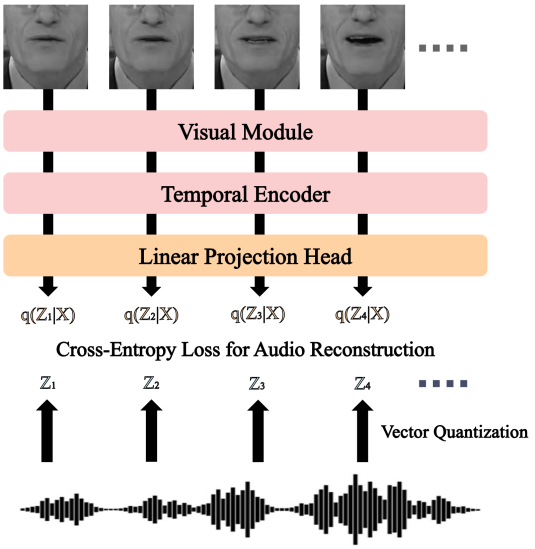


Figure 2: *Overview of the SyncVSR training framework. Given a sequence of video frames, the encoder generates a corresponding sequence of quantized audio tokens in a non-autoregressive manner. z_t denotes audio tokens, and $q(z_t|x)$ is the encoder's prediction through a linear projection layer.*

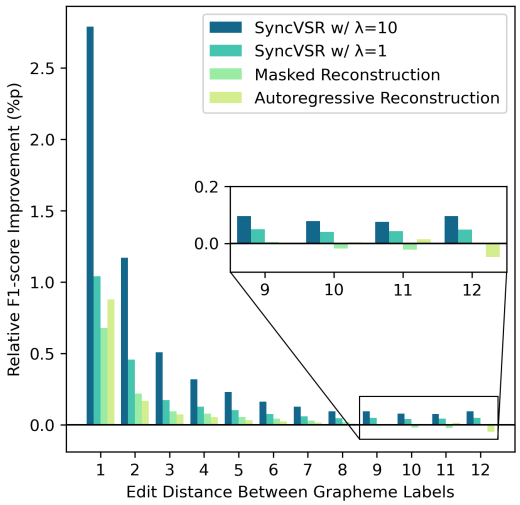


Figure 3: *The edit distance of word pairs and the model's discriminative ability. Homophene pairs resemble each other closely in graphemes, a scenario where SyncVSR shows better classification performance over the vanilla setting trained without audio information. Non-autoregressive generation with strong audio reconstruction loss weight (λ) is optimal, whereas masked reconstruction could cause harm in certain instances.*

chronization. By exploiting the discrete nature of quantized audio for frame-level supervision, SyncVSR circumvents the limitations of previous methods that rely on indirect semantic alignment or utilize potentially biased handcrafted features. This allows our model to discern fine-grained phonetic differences inherent in homophenes, enhancing the model's interpretative fidelity and data efficiency. Our empirical results highlight the efficacy of SyncVSR, which establishes new benchmark results across a range of VSR tasks.

2. Methodology

Crossmodal Audio Token Synchronization. Our work integrates audio reconstruction loss with VSR training objectives. Conventionally, word-level VSR employs a word classification loss, whereas sentence-level VSR utilizes the joint CTC-Attention loss [21]. The total loss is the weighted sum of the task-specific objective loss and our audio reconstruction loss.

Let $\mathcal{D}$ be a training set that consists of a training sample (x, y, z). Let x and y be the input video and ground truth label. Let $z = \{z_t\}_{t \leq T}$ be a discrete audio sequence corresponding to the input video x.

Word Classification Loss. For word-level VSR, cross-entropy loss measures the difference between predicted class probabilities and the ground truth labels. Given that y represents the ground truth category for the input video x, the objective loss is formulated as follows:

$$\mathcal{L}_{\text{task}} = -\mathbb{E}_{(x,y,z)\in\mathcal{D}} \log p(y|x)$$

where $p(y|x)$ denotes the output probability from the model.

Joint CTC-Attention Loss. For sentence-level VSR, we employ a combination of Connectionist Temporal Classification (CTC) [22] loss from the encoder and Language Modeling (LM) loss from the decoder, known as joint CTC-Attention loss.

Let $\pi = \{\pi_t\}_{t \leq T}$ be intermediate CTC labels, and $\phi(y)$ be a set of all possible intermediate labels for CTC loss. Using p_{LM} for language modeling and p_{CTC} for conditional independent prediction, we define the losses as follows:

$$\mathcal{L}_{\text{LM}} = -\mathbb{E}_{(x,y,z)\in\mathcal{D}} \sum_{t \leq T} \log p_{\text{LM}}(y_t|x, y_{<t}),$$

$$\mathcal{L}_{\text{CTC}} = -\mathbb{E}_{(x,y,z)\in\mathcal{D}} \sum_{\pi\in\phi(y)} \sum_{t \leq T} \log p_{\text{CTC}}(\pi_t|x),$$

and the final objective loss is defined as a combination of $\mathcal{L}_{\text{LM}}$ and $\mathcal{L}_{\text{CTC}}$, i.e.,

$$\mathcal{L}_{\text{task}} = \alpha\mathcal{L}_{\text{CTC}} + (1-\alpha)\mathcal{L}_{\text{LM}}$$

where α is a hyperparameter with the constraint $0 \leq \alpha \leq 1$.

Audio Reconstruction Loss. The synchronization between audio and video in our framework is designed to align each video frame with a corresponding number of audio tokens. This alignment is based on audio (16kHz) and video (25fps) sampling rates, with a specific hop size ensuring a coherent match between video frame rate and audio units. We use a ratio of one video frame to four vector quantized audio tokens (100Hz).

To make the model generate discrete audio tokens, we use cross-entropy loss to predict the quantized audio z_t from the input video frames. Let $q(z_t|x)$ be an output of the model at time t from the video x. The audio reconstruction loss is as follows:

$$\mathcal{L}_{\text{Sync}} = \mathbb{E}_{(x,y,z)\in\mathcal{D}} \left[-\frac{1}{T} \sum_{t \leq T} \log q(z_t|x) \right]$$

where T denotes the number of video frames.

Using these, we simply design the total loss as below, with the hyperparameter λ as the weight for the audio reconstruction loss,

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda\mathcal{L}_{\text{Sync}}.$$

Table 1: *Video-based VSR performance on word-level tasks. Evaluations were done on Lip Reading in the Wild (LRW) [23] English benchmark and CAS-VSR-W1K [24] Chinese benchmark. WB implies the usage of word boundary, which is an indicator for the target word's appearance. Our metrics are averaged across three experiments, with subscripts notating the standard deviation. Transcription Alignment is our experiment using the method of aligning character-level pseudo-labels from ASR models instead of auditory data.*

Method	Temporal Model	Video Hours	Top-1 Acc. (%) ↑		
			LRW	LRW(WB)	CAS-VSR-W1K
Born-Again [25]	MS-TCN	157	87.9	-	46.6
LiRA [14]	Conformer	590	88.1	-	-
WPCL + APFF [26]	MS-TCN	157	88.3	-	-
Ma *et al*. [27]	DC-TCN	157	88.4	-	43.7
Feng *et al*. [28]	BiGRU	157	86.2	88.4	<u>55.7</u>
MVM [8]	MS-TCN	157	88.5	-	53.8
NetVLAD [29]	MS-TCN	157	89.4	-	-
Koumparoulis *et al*. [30]	Transformer	157	89.5	-	-
Training Strategy [31]	DC-TCN	157	90.4	92.1	-
MTLAM [12]	DC-TCN	157	<u>91.7</u>	-	54.3
Training Strategy + LiRA	DC-TCN	595	-	92.3	-
Training Strategy + CM-Aux [11]	DC-TCN	1,459	-	<u>92.9</u>	-
Transcription Alignment	Transformer	157	93.1	94.8	-
SyncVSR	Transformer	157	**93.2** ± 0.1	**95.0** ± 0.0	**58.2** ± 0.0

Table 2: *Landmark-based VSR evaluated on a word-level task. The methods below were trained from scratch on the LRW.*

Method	Input Type	#Params	Top-1 ↑
Lip Graph Assisted [32]	Graph	30M	49.3
Adaptive GCN [33]	Graph	45M	60.7
Another Point of View [34]	Pointcloud	12M	<u>62.7</u>
SyncVSR	Pointcloud	11M	**75.1** ± 0.1
SyncVSR(WB)	Pointcloud	11M	**80.3** ± 0.0

3. Experimental Setup

Training Dataset. We employ the LRW [23] dataset for English and the CAS-VSR-W1K [24] for Chinese to evaluate word-level VSR tasks. The LRW dataset comprises 500 words, each represented by up to 1,000 training videos. The LRW-1000 dataset consists of 718,018 videos spanning 1,000 words. Our sentence-level experimental framework was anchored on the LRS2 [35] and LRS3 [6] datasets, representing the most extensive publicly available resources for audio-visual speech recognition in English. The LRS2 dataset, sourced from BBC programs, comprises 144,482 video clips, totaling 225 hours of video content. The LRS3 dataset, harvested from TED talks, encompasses 151,819 video clips, amassing 439 hours of footage. Additional training data was sourced from the English-speaking segments of the VoxCeleb2 [43] dataset, comprised of a training corpus totaling 1,323 video hours, complemented by transcriptions following the scheme of AutoAVSR [38].

Dataset Preprocessing. We used MediaPipe [44] to identify the region of interest with a size of 128 x 128 for video-based VSR, and the extracted landmark data served as input for a pointcloud-based VSR system. We used a data augmentation scheme of a resized random crop with a size of (96, 96) and a random horizontal flip and applied a center crop for inference similar to that of previous works [3, 38].

Model Architecture. For word-level VSR, an encoder is composed of a combination of 3D CNN, ResNet18, and Transformer [45] to extract video features following the previous works [8, 12, 13, 31]. On the other hand, Conformer [46] is used as a temporal backbone for sentence-level VSR, where we follow the model size and configuration of previous works' settings [14, 17, 37].

Training Recipe. For word-level VSR tasks, we train the model for 200 epochs with the Adam [47] optimizer. The learning rate increased from 0 to 0.0001 for the first 5 epochs and then decreased linearly. In the case of sentence-level VSR, the model is trained for 100 epochs with the Adam optimizer, where the learning rate linearly decays from the peak of 0.001 at the 3rd epoch. Batch size is 384 for word-level and 64 for sentence-level, distributed to 4×A100 GPUs. The rest of the training specifics follow the previous work's settings from [38]. Our metrics were obtained from the average of three random seeds.

4. Results

Versatility Across Tasks, Languages, and Modalities. Our framework is comprehensively evaluated according to tasks, languages, and input modalities. In word-level tasks, shown in Table 1, SyncVSR marks state-of-the-art results in English and Chinese benchmarks. In sentence-level tasks, displayed in Figure 1 and Table 3, SyncVSR outperforms available methodologies when given a similar amount of video dataset. Notably, our method also advances a tier in model size, where our base-size model shows superior performance over other large-size models. Our method also achieves state-of-the-art performance in landmark-based VSR tasks shown in Table 2 and Table 4.

Distinguishing Homophenes. Homophenes often closely resemble each other in their graphemes—the smallest functional units of a writing system. For example, homophene pairs, like (*Million*, *Billion*) or (*Living*, *Giving*), differ by just one grapheme. Although earlier research, notably by Kim *et al*. [8] has examined a subset of these pairs, a full-scale evaluation of every potential homophene pair has yet to be achieved. As a result, in Figure 3, we assess the relative F1-score gain of existing training methods over a vanilla setting that does not utilize the audio data, focusing on the grapheme edit distances. This suggests that the inclusion of an audio reconstruction loss objective assists in differentiating visemes that are mapped into similar graphemes, which is where homophene pairs are typically found.

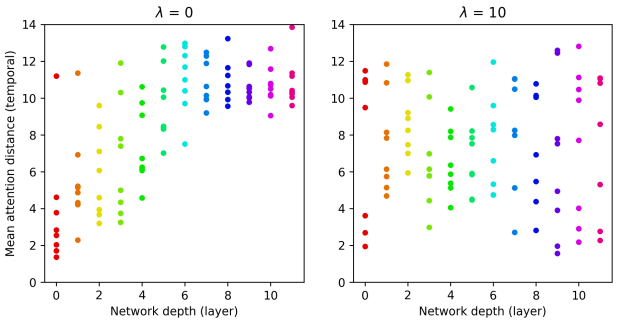
Table 3: *Video-based VSR performance on sentence-level tasks grouped with video data resource usage. Evaluations were done on LRS2 [35] and LRS3 [6] benchmarks. LM indicates whether an external language model is used. The methods listed below use the base-size model unless specified as large-size. Reported scores have a standard deviation smaller than 0.5.*

Method	Video Hours	LM	WER ↓	
			LRS2	LRS3
Less than 500h				
TDNN [36]	223	✓	48.9	-
CM-Seq2Seq [37]	223/438	✓	39.1	46.9
CM-Aux [11]	223/438	✓	32.9	37.9
RAVEn [17]	438	✓	-	39.1
AutoAVSR [38]	438	✓	-	36.3
SyncVSR (Ours)	223/438	✗	**30.7**	**33.3**
SyncVSR (Ours)	223/438	✓	**28.9**	**31.2**
Less than 1000h				
KD + CTC [39]	995	✓	51.3	59.8
KD-Seq2Seq [7]	818	✗	49.2	59.0
MVM [8]	818	✗	44.5	-
LiRA [14]	661	✓	38.8	-
RAVEn	661	✓	32.1	-
VTP [40]	698	✓	28.9	40.6
AutoAVSR [38]	818	✓	27.9	33.0
CM-Aux	818	✓	27.3	34.7
SyncVSR (Ours)	661	✗	**22.0**	**30.4**
SyncVSR (Ours)	661	✓	**20.0**	**28.1**
Less than 2000h				
TM-Seq2Seq [10]	1,391	✓	48.3	58.9
CM-Aux	1,459	✓	25.5	31.5
AV-HuBERT [15]	1,992/1,759	✗	31.2	34.8
AV-HuBERT-Large	1,992/1,759	✗	25.5	26.9
VATLM [18]	1,992/1,759	✗	30.6	34.2
VATLM-Large	1,992/1,759	✗	24.3	26.2
LMDecoder [16]	1,992	✗	23.8	-
RAVEn	1,992/1,759	✗	-	33.1
RAVEn-Large	1,992/1,759	✗	19.3	24.4
RAVEn-Large	1,992/1,759	✓	17.9	23.1
SyncVSR (Ours)	1,992	✗	18.5	23.4
SyncVSR (Ours)	1,992	✓	**16.5**	**21.5**
Greater than 2000h				
VTP	2,676	✓	22.6	30.7
AutoAVSR	3,448	✓	**14.6**	19.1
ViT 3D [41]	90,000	✗	-	17.0
LP Conformer [42]	100,000	✗	-	**12.8**

Significance of Full Sequence Synchronization. Furthermore, Figure 3 compares our reconstruction method with masked reconstruction. Not only does our full-length non-autoregressive generation excel when graphemes are similar, but it is also noteworthy that masked reconstruction could cause harm for classifying pairs with far edit distances. This implies previous works based on masked reconstruction [15–18] might be imperfect for aligning visual and audio modalities. In contrast, our audio reconstruction objective can be amplified up to ten times above the original task objective, which assists in discerning fine-grained visemes without causing any obstruction. Types of crossmodal indicators, whether they be quantized acoustic tokens from vq-wav2vec [49] or character-level transcriptions from wav2vec2 [50], are trivial, as seen in Table 1.

Table 4: *Impact of CTC loss and audio reconstruction loss on the LRS2 benchmark. To the best of our knowledge, this is the first instance of reporting a successful implementation of landmark-based sentence-level visual speech recognition.*

Sync	CTC	WER ↓		Perplexity ↓	
		Video	Pointcloud	Video	Pointcloud
✗	✗	45.9	99.9	4.0	40.7
✗	✓	43.2	99.8	4.1	40.4
✓	✗	38.1	77.4	2.8	8.1
✓	✓	**30.7**	**74.6**	**2.7**	**7.7**



(a) Trained w/o Sync (b) Trained w/ Sync

Figure 4: *Influence of audio reconstruction loss weight (λ) on the encoder's representation visualized with the mean attention distance [48] distribution. Each point indicates the weighted distance of attention from the query frame to other frames.*

Enhancing Speech Representation Learning. For sentence-level VSR, joint CTC-Attention loss is widely used in previous works [11, 16, 17, 37–39]. Despite its benefit in the decoding stage, its utility in terms of representation learning remains uncertain. According to Table 4, CTC loss marginally contributes to the perplexity of the model, whereas our audio reconstruction loss term strongly improves learning in both pointcloud and video modalities. The effect of frame-level crossmodal supervision is illustrated in Figure 4, where inner representation indicates that the temporal encoder model exhibits a change of bias towards local neighboring frames.

5. Conclusion

We addressed the problem of homophenes with an improved crossmodal synchronization method, effectively bridging the divide between visual cues and their corresponding audio segments. The use of quantized audio tokens for direct frame-level supervision enables SyncVSR to achieve state-of-the-art performance on various benchmarks with a remarkable level of data efficiency. We believe SyncVSR is a step toward future developments in the field of multimodal speech recognition.

6. Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.RS-2019-II190075 Artificial Intelligence Graduate School Program (KAIST), Cloud TPUs from Google's TPU Research Cloud (TRC), and an Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean Government (24ZB1100, Core Technology Research for Self-improving Integrated AI Systems).

7. References

[1] H. Laux, A. Hallawa, J. C. S. Assis *et al.*, "Two-stage visual speech recognition for intensive care patients," *Scientific Reports*, vol. 13, 2023.

[2] N. Tye-Murray, M. Sommers, B. Spehar *et al.*, "Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing." *Ear and hearing*, vol. 28 5, pp. 656–68, 2006.

[3] B. Martínez, P. Ma, M. Pantic *et al.*, "Lipreading using temporal convolutional networks," in *ICASSP*, 2020, pp. 6319–6323.

[4] B. Xu, C. Lu, Y. Guo *et al.*, "Discriminative multi-modality speech recognition," *CVPR*, pp. 14 421–14 430, 2020.

[5] A. Haliassos, K. Vougioukas, S. Petridis *et al.*, "Lips don't lie: A generalisable and robust approach to face forgery detection," *CVPR*, pp. 5037–5047, 2020.

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

[7] S. Ren, Y. Du, J. Lv, G. Han *et al.*, "Learning from the master: Distilling cross-modal advanced knowledge for lip reading," in *CVPR*, 2021, pp. 13 325–13 333.

[8] M. Kim, J. H. Yeo, and Y. M. Ro, "Distinguishing homophenes using multi-head visual-audio memory for lip reading," in *AAAI*, 2022, pp. 1174–1182.

[9] Y. Zhao, R. Xu, X. Wang, P. Hou, H. Tang, and M. Song, "Hearing lips: Improving lip reading by distilling speech recognizers," in *AAAI*, vol. 34, no. 04, 2020, pp. 6917–6924.

[10] T. Afouras, J. S. Chung, A. Senior *et al.*, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 8717–8727, 2018.

[11] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," *Nature Machine Intelligence*, pp. 1–10, 2022.

[12] J. H. Yeo, M. Kim, and Y. M. Ro, "Multi-temporal lip-audio memory for visual speech recognition," in *ICASSP*, 2023, pp. 1–5.

[13] M. Kim, J. Hong, S. J. Park, and Y. M. Ro, "Multi-modality associative bridging through memory: Speech sound recollected from face video," in *ICCV*, 2021.

[14] P. Ma, R. Mira, S. Petridis, B. W. Schuller, and M. Pantic, "Lira: Learning visual speech representations from audio through self-supervision," in *INTERSPEECH*, 2021, pp. 3011–3015.

[15] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *ICLR*, 2022.

[16] M. Kim, J. H. Yeo, J. Choi, and Y. M. Ro, "Lip reading for low-resource languages by learning and combining general speech knowledge and language-specific knowledge," in *ICCV*, 2023.

[17] A. Haliassos, P. Ma, R. Mira *et al.*, "Jointly learning visual and auditory speech representations from raw data," in *ICLR*, 2023.

[18] Q. Zhu, L. Zhou, Z. Zhang *et al.*, "Vatlm: Visual-audio-text pre-training with unified masked prediction for speech representation learning," *ArXiv*, vol. abs/2211.11275, 2022.

[19] P. He, J. Gao, and W. Chen, "Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing," *ArXiv*, vol. abs/2111.09543, 2021.

[20] K. Clark, M.-T. Luong, Q. V. Le *et al.*, "Electra: Pre-training text encoders as discriminators rather than generators," *ICLR*, 2020.

[21] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *ICASSP*, 2017, pp. 4835–4839.

[22] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.

[23] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading in the wild," in *ACCV*. Springer, 2017, pp. 87–103.

[24] S. Yang, Y. Zhang, D. Feng *et al.*, "Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," *FG*, pp. 1–8, 2018.

[25] P. Ma, B. Martinez, S. Petridis, and M. Pantic, "Towards practical lipreading with distilled and efficient models," *ICASSP*, 2020.

[26] W. Tian, H. Zhang, C. Peng, and Z.-Q. Zhao, "Lipreading model based on whole-part collaborative learning," *ICASSP*, 2022.

[27] P. Ma, Y. Wang, J. Shen, S. Petridis, and M. Pantic, "Lip-reading with densely connected temporal convolutional networks," *WACV*, pp. 2856–2865, 2020.

[28] D. Feng, S. Yang, S. Shan, and X. Chen, "Learn an effective lip reading model without pains," *arXiv preprint arXiv:2011.07557*, 2020.

[29] H. Yang, T. Luo, Y. Zhang *et al.*, "Improved word-level lipreading with temporal shrinkage network and netvlad," *ICMI*, 2022.

[30] A. Koumparoulis and G. Potamianos, "Accurate and resource-efficient lipreading with efficientnetv2 and transformers," in *ICASSP*, 2022, pp. 8467–8471.

[31] P. Ma, Y. Wang, S. Petridis *et al.*, "Training strategies for improved lip-reading," in *ICASSP*, 2022, pp. 8472–8476.

[32] H. Liu, Z. Chen, and B. Yang, "Lip graph assisted audio-visual speech recognition using bidirectional synchronous fusion," in *INTERSPEECH*, 2020.

[33] X. Z. Changchong Sheng, H. Xu, M. Pietikäinen, and L. Liu, "Adaptive semantic-spatio-temporal graph convolutional network for lip reading," *IEEE Transactions on Multimedia*, vol. 24, 2022.

[34] B. Pouthier, L. Pilati, and G. Valenti, "Another point of view on visual speech recognition," *INTERSPEECH*, 2023.

[35] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *CVPR*, 2017, pp. 3444–3453.

[36] J. Yu, S.-X. Zhang, J. Wu *et al.*, "Audio-visual recognition of overlapped speech for the lrs2 dataset," *ICASSP*, pp. 6984–6988, 2020.

[37] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *ICASSP*, 2021, pp. 7613–7617.

[38] P. Ma, A. Haliassos, A. Fernandez-Lopez *et al.*, "Auto-avsr: Audio-visual speech recognition with automatic labels," in *ICASSP*, 2023, pp. 1–5.

[39] T. Afouras, J. S. Chung, and A. Zisserman, "Asr is all you need: Cross-modal distillation for lip reading," *ICASSP*, 2019.

[40] K. R. Prajwal, T. Afouras, and A. Zisserman, "Sub-word level lip reading with visual attention," in *CVPR*, 2022, pp. 5162–5172.

[41] D. Serdyuk, O. Braga, and O. Siohan, "Transformer-based video front-ends for audio-visual speech recognition," in *INTERSPEECH*, 2022.

[42] O. Chang, H. Liao, D. Serdyuk, A. Shah, and O. Siohan, "Conformers are all you need for visual speech recogntion," *ICASSP*, 2024.

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

[44] C. Lugaresi, J. Tang, H. Nash *et al.*, "Mediapipe: A framework for perceiving and processing reality," in *CVPR*, vol. 2019.

[45] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[46] A. Gulati, J. Qin, C.-C. Chiu *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *INTERSPEECH*, 2020, pp. 5036–5040.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, 2014.

[48] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[49] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *ICLR*, 2020.

[50] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.