Cross-Modal Global Interaction and Local Alignment for Audio-Visual Speech Recognition

Yuchen Hu[1], **Ruizhe Li**[2], **Chen Chen**[1], **Heqing Zou**[1], **Qiushi Zhu**[3] and **Eng Siong Chng**[1]
[1]Nanyang Technological University, Singapore
[2]University of Aberdeen, UK
[3]University of Science and Technology of China, China
{yuchen005@e., chen1436@e., heqing001@e., aseschng@}ntu.edu.sg, ruizhe.li@abdn.ac.uk, qszhu@mail.ustc.edu.cn

Abstract

Audio-visual speech recognition (AVSR) research has gained a great success recently by improving the noise-robustness of audio-only automatic speech recognition (ASR) with noise-invariant visual information. However, most existing AVSR approaches simply fuse the audio and visual features by concatenation, without explicit interactions to capture the deep correlations between them, which results in sub-optimal multimodal representations for downstream speech recognition task. In this paper, we propose a cross-modal global interaction and local alignment (GILA) approach for AVSR, which captures the deep audio-visual (A-V) correlations from both global and local perspectives. Specifically, we design a global interaction model to capture the A-V complementary relationship on modality level, as well as a local alignment approach to model the A-V temporal consistency on frame level. Such a holistic view of cross-modal correlations enable better multimodal representations for AVSR. Experiments on public benchmarks LRS3 and LRS2 show that our GILA outperforms the supervised learning state-of-the-art. Code is at https://github.com/YUCHEN005/GILA.

1 Introduction

With recent advancement of deep learning techniques, automatic speech recognition (ASR) has achieved quite good performance [Graves, 2012; Vaswani *et al.*, 2017]. However, ASR systems are usually vulnerable to noise and would degrade significantly when put under noisy conditions [Sumby and Pollack, 1954]. To improve their performance under various scenarios, recent works on noise-robust speech recognition have made some progress [Wang *et al.*, 2020].

A currently popular research direction on robustness combines audio (A) and visual (V) features to benefit from the noise-invariant lip movement information. With use of two modalities, audio-visual speech recognition (AVSR) systems move one step closer to how human perceives speech [Sumby and Pollack, 1954] and achieve better performance in many application scenarios [Biswas *et al.*, 2016; Koguchi *et al.*, 2018]. Thanks to recent advance of neural network, AVSR has achieved a remarkable success [Afouras *et al.*, 2018a; Makino *et al.*, 2019; Xu *et al.*, 2020; Ma *et al.*, 2021; Pan *et al.*, 2022; Shi *et al.*, 2022a; Shi *et al.*, 2022b; Hsu and Shi, 2022]. However, most existing AVSR works simply employ feature concatenation for audio-visual (A-V) fusion, without explicit interactions to capture deep correlations between them [Raij *et al.*, 2000]: 1) From global perspective, they may not capture the complementary relationship between A-V modalities. Such relationship means when one modality is missing or corrupted, the other modality can supply valid information for downstream task [Wang *et al.*, 2022]. Failure to capture it would make the system confused about the significance of each modality and thus degrade the performance [Hori *et al.*, 2017; Tao and Busso, 2018]. 2) From local perspective, they may ignore the temporal alignment between A-V frames, which could be a problem due to the ambiguity of homophenes [Kim *et al.*, 2022] where same lip shape could produce different sounds. Such misalignment between lip and audio sequences would increase the difficulty of efficient multimodal fusion and affect final performance [Tsai *et al.*, 2019; Lv *et al.*, 2021].

To capture the global complementary relationship between different modalities, cross-attention has been widely investigated in recent multimodal studies to learn the inter-modal correspondence [Lee *et al.*, 2020; Li *et al.*, 2021; Goncalves and Busso, 2022; Mercea *et al.*, 2022]. Despite the effectiveness, it fails to simultaneously preserve the intra-modal correspondence that could adaptively select the information of each individual modality for the inter-modal correspondence modeling [Wang *et al.*, 2022], which thus results in sub-optimal complementary relationship between modalities.

From the local perspective, contrastive learning has been popular for cross-modal temporal alignment to model the frame-level consistency [Korbar *et al.*, 2018; Morgado *et al.*, 2021; Yang *et al.*, 2022], but they seem to only align the multimodal features within same model layer, ignoring the alignment across different layers. Since different-layer features contain semantic representations of different granularities [Gu *et al.*, 2021], we argue that the alignment between them could capture extra contextual information to improve the modeled temporal consistency.

In this paper, we propose a cross-modal global interaction and local alignment (GILA) approach to effectively capture

the deep audio-visual correlations from both global and local perspectives. Specifically, we propose an attention-based global interaction (GI) model to capture the A-V complementary relationship on modality level. On top of the vanilla cross-attention, we propose a novel iterative refinement module to jointly model the A-V inter- and intra-modal correspondence. It could adaptively leverage the information within each individual modality to capture the inter-modal correspondence, which thus results in better complementary relationship between A-V modalities. With global knowledge of A-V correlations, the system may still be less aware of the local details. To this end, we further design a cross-modal local alignment (LA) approach via contrastive learning to model the A-V temporal consistency on frame level. Based on the vanilla within-layer alignment, we propose a novel cross-layer contrastive learning approach to align A-V features across different GI model layers. Such design could capture extra contextual information between the different-granularity semantic representations, which enables more informative temporal consistency between A-V frames. As a result, our proposed GILA can capture deep holistic correlations between A-V features and finally generate better multimodal representations for downstream recognition task.

To the best of our knowledge, this is the first AVSR work to model deep A-V correlations from both global and local perspectives. Our main contributions are summarized as:

- We present GILA, a novel approach to capture deep audio-visual correlations for AVSR task, from both global and local perspectives.
- We propose a cross-modal global interaction (GI) model to capture A-V complementary relationship on modality level, as well as a local alignment (LA) approach to model the A-V temporal consistency on frame level.
- Experimental results on two public benchmarks demonstrate the effectiveness of our GILA against the state-of-the-art (SOTA) supervised learning baseline, with up to 16.2% relative WER improvement.

2 Related Work

Audio-Visual Speech Recognition. Most existing AVSR works focus on novel architectures and supervised learning methods, investigating how to effectively model and fuse the audio-visual modalities. TM-seq2seq [Afouras *et al.*, 2018a] proposes a Transformer-based [Vaswani *et al.*, 2017] AVSR system with sequence-to-sequence loss. Hyb-RNN [Petridis *et al.*, 2018] proposes a RNN-based AVSR system with hybrid seq2seq/CTC loss [Watanabe *et al.*, 2017]. RNN-T [Makino *et al.*, 2019] employs recurrent neural network transducer [Graves, 2012] for AVSR task. EG-seq2seq [Xu *et al.*, 2020] builds a joint audio enhancement and multimodal speech recognition system based on RNN. LF-MMI TDNN [Yu *et al.*, 2020] proposes a joint audio-visual speech separation and recognition system based on TDNN. Hyb-Conformer [Ma *et al.*, 2021] proposes a Conformer-based [Gulati *et al.*, 2020] AVSR system with hybrid seq2seq/CTC loss, where the audio-visual streams are encoded separately and then concatenated for decoding, which has achieved the supervised learning SOTA on both LRS3 and LRS2 datasets. MoCo+wav2vec [Pan *et al.*, 2022] employs self-supervised pre-trained audio/visual front-ends to improve AVSR performance, which has achieved the SOTA on LRS2 dataset. However, these studies simply concatenate the audio and visual features for multimodal fusion, without explicit interactions to capture their deep correlations. Recently proposed AV-HuBERT [Shi *et al.*, 2022a; Shi *et al.*, 2022b] employs self-supervised learning to capture contextual correlations between audio-visual features, and the latest u-HuBERT [Hsu and Shi, 2022] extends it to a unified framework of multimodal and unimodal pre-training, which has achieved the SOTA on LRS3 dataset. However, they require a large amount of unlabeled data and computing resources. In this work, we propose a novel supervised learning approach called GILA to efficiently capture deep A-V correlations from both global and local perspectives.

Cross-Modal Modality-Level Interaction. Attention methods have been widely investigated to interact between different modalities to capture their complementary relationship, in various multimodal applications such as A-V emotion recognition [Goncalves and Busso, 2022], A-V action localization [Lee *et al.*, 2020], etc. Recent works employ cross-attention to enable extracted features of different modalities to attend to each other [Lee *et al.*, 2020; Li *et al.*, 2021; Goncalves and Busso, 2022; Mercea *et al.*, 2022], which is found effective to capture the inter-modal correspondence and significantly improves the system performance. However, they may not simultaneously preserve the intra-modal correspondence that could adaptively select the unimodal information for inter-modal correspondence modeling [Wang *et al.*, 2022]. To this end, we propose a novel iterative refinement module to jointly model the inter- and intra-modal correspondence, where the key idea is introducing a bottleneck feature to recurrently collect multimodal information.

Cross-Modal Frame-Level Alignment. Cross-modal alignment aims to model the temporal consistency between sequences of different modalities, and alleviate the frame-level misalignment problem in some scenarios [Tsai *et al.*, 2019; Lv *et al.*, 2021; Kim *et al.*, 2022]. This is typically done by contrastive learning where the correspondence between positive pairs is trained to be stronger than those of negative pairs [Chopra *et al.*, 2005]. Recently, contrastive learning is popular for cross-modal temporal alignment, which has achieved significant improvement on various tasks [Korbar *et al.*, 2018; Morgado *et al.*, 2021; Yang *et al.*, 2022]. However, they seem to only align features of multiple modalities within same model layer, ignoring the alignment across different layers that could learn extra contextual information between different-granularity semantic representations. In this work, we propose a cross-layer contrastive learning approach for holistic A-V alignments.

3 Methodology

In this part, we first introduce the overall architecture of proposed GILA in Section 3.1. Then, we describe its two main components, *i.e.*, the cross-modal global interaction model in Section 3.2 and local alignment approach in Section 3.3. Finally, we explain the training objective in Section 3.4.

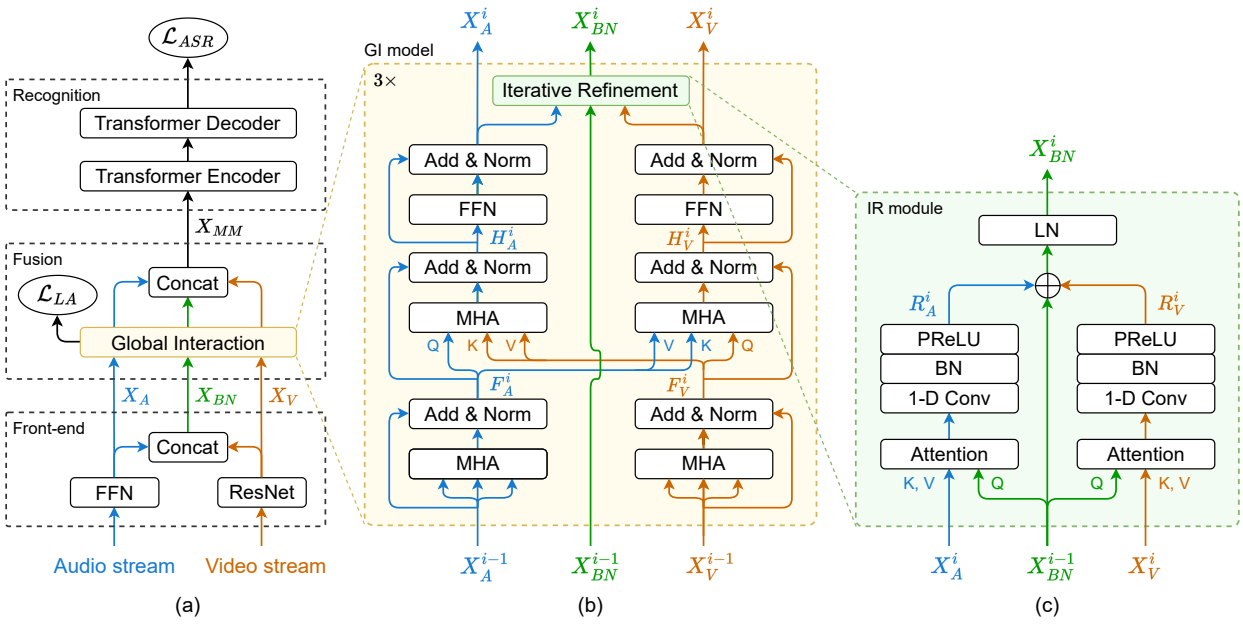

Figure 1: Block diagrams of proposed GILA: (a) Overall architecture, (b) Global Interaction (GI) model, (c) Iterative Refinement module. The $\mathcal{L}_{ASR}$ denotes speech recognition loss, and $\mathcal{L}_{LA}$ denotes local alignment loss.

3.1 Overall Architecture

As illustrated in Figure 1(a), the proposed GILA system consists of front-end module, fusion module and recognition module. We first introduce a front-end module to pre-process the synchronized audio-video input streams, which employs a linear projection layer for audio front-end and a modified ResNet-18 [Shi *et al.*, 2022a] for visual front-end. We also concatenate the processed A-V features to build a bottleneck feature X_{BN} to collect multimodal information. Then, we propose a fusion module for audio-visual fusion. Specifically, we propose a global interaction model and a local alignment approach to capture deep A-V correlations. The resulted audio, visual and bottleneck features are then concatenated to generate the multimodal feature X_{MM}. Finally, we introduce a Transformer-based recognition module to encode the multimodal feature and predict the output tokens. The overall training objective consists of the speech recognition loss $\mathcal{L}_{ASR}$ and the local alignment loss $\mathcal{L}_{LA}$.

3.2 Cross-Modal Global Interaction (GI)

As shown in Figure 1(b), we propose a cross-modal global interaction model to capture the complementary relationship between A-V modalities. Specifically, we first introduce cross-attention to interact audio-visual features to capture inter-modal correspondence. On top of that, we further propose a novel iterative refinement (IR) module to jointly model the inter- and intra-modal correspondence, aiming to better capture the complementary relationship on modality level.

Cross-Attention aims to capture the A-V inter-modal correspondence. As illustrated in Figure 1(b), the input audio-visual features of i-th GI model layer (*i.e.*, X_A^{i-1}, X_V^{i-1}, $i \in \{1, 2, 3\}$) are first sent into two separate self-attention modules [Vaswani *et al.*, 2017] for modeling, which generates two intermediate features, F_A^i and F_V^i:

$$\begin{aligned} F_A^i &= LN(X_A^{i-1} + MHA(X_A^{i-1}, X_A^{i-1}, X_A^{i-1})), \\ F_V^i &= LN(X_V^{i-1} + MHA(X_V^{i-1}, X_V^{i-1}, X_V^{i-1})), \end{aligned} \tag{1}$$

where "LN" denotes layer normalization [Ba *et al.*, 2016], "MHA" denotes multi-head scaled dot-product attention [Vaswani *et al.*, 2017].

Then, we introduce cross-attention to enable audio-visual features to attend to each other for complementation, in order to capture the inter-modal correspondence:

$$\begin{aligned} H_A^i &= LN(F_A^i + MHA(F_A^i, F_V^i, F_V^i)), \\ H_V^i &= LN(F_V^i + MHA(F_V^i, F_A^i, F_A^i)), \end{aligned} \tag{2}$$

After that, we utilize position-wise feed-forward network (FFN) [Vaswani *et al.*, 2017] to generate outputs:

$$\begin{aligned} X_A^i &= LN(H_A^i + FFN(H_A^i)), \\ X_V^i &= LN(H_V^i + FFN(H_V^i)), \end{aligned} \tag{3}$$

where FFN consists of two linear layers with a ReLU [Glorot *et al.*, 2011] activation in between.

Iterative Refinement (IR) aims to jointly model the A-V inter- and intra-modal correspondence, where the bottleneck feature plays a key role. As shown in Figure 1(c), the input bottleneck feature X_{BN}^{i-1} first attends to the A/V feature from cross-attention (*i.e.*, X_A^i, X_V^i) respectively, followed by convolution to generate two residual features R_A^i and R_V^i:

$$\begin{aligned} R_A^i &= \text{Conv}(\text{Attention}(X_{BN}^{i-1}, X_A^i, X_A^i)), \\ R_V^i &= \text{Conv}(\text{Attention}(X_{BN}^{i-1}, X_V^i, X_V^i)), \end{aligned} \tag{4}$$

where "Conv" denotes a 1×1 convolution layer followed by batch normalization (BN) [Ioffe and Szegedy, 2015] and parametric ReLU (PReLU) activation.

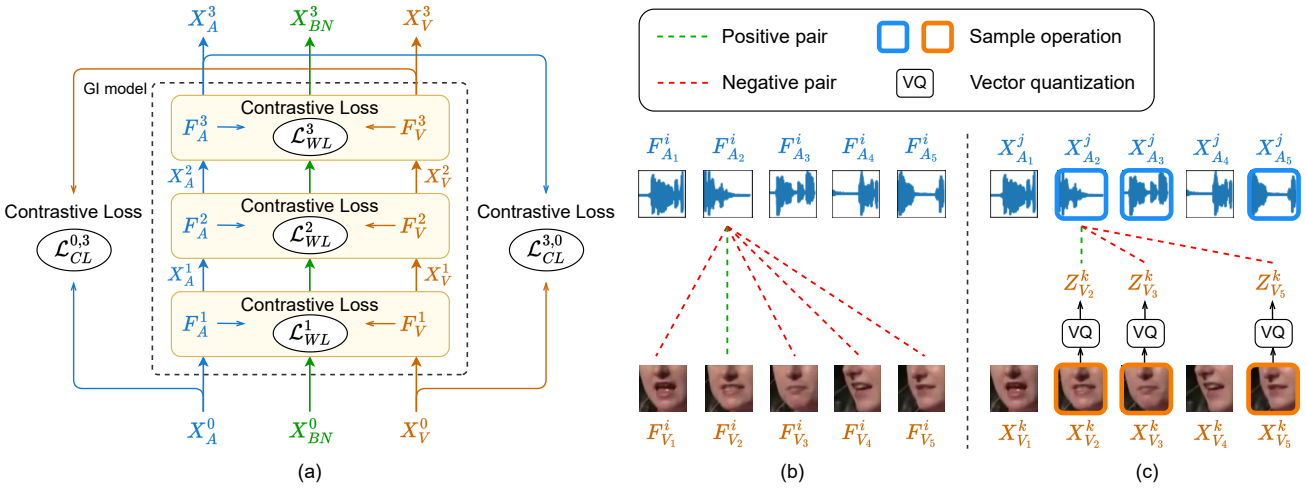

Figure 2: Block diagrams of proposed cross-modal local alignment approach: (a) Overview, (b) Within-Layer (WL) contrastive learning, (c) Cross-Layer (CL) contrastive learning.

The attention blocks aim to build interactions between individual audio/visual feature and the bottleneck feature that contains multimodal information. Therefore, the individual A/V modality can not only attend to the other modality, but also attend to itself simultaneously. As a result, we can jointly model the inter- and intra-modal correspondence, which helps extract the useful information in A/V modality.

Finally, we add the two generated residual features to input bottleneck feature, in order to refine more informative multimodal representations:

$$X_{BN}^{i} = LN(X_{BN}^{i-1} + R_A^i + R_V^i), \tag{5}$$

With increasing multimodal information in the bottleneck feature, the IR module in next GI model layer can better capture the A-V correspondences by Equation 4, and so on. Such refining mechanism enables IR module to effectively model the inter- and intra-modal correspondence.

3.3 Cross-Modal Local Alignment (LA)

In order to learn more local details of A-V correlations, we further propose a cross-modal local alignment approach to model the temporal consistency between A-V frames, as presented in Figure 2. Specifically, we first introduce within-layer contrastive learning to align the A-V features within same GI model layer. Based on that, we propose a novel cross-layer contrastive learning method for A-V alignment across different GI model layers, aiming to learn more informative A-V temporal consistency on frame level.

Within-Layer (WL) Contrastive Learning aims to align the A-V features within same GI model layer. As illustrated by Figure 2(a)(b), we select the i-th layer's intermediate features F_A^i and F_V^i for alignment. Denote that $F_A^i = \{F_{A_t}^i|_{t=1}^T\}, F_V^i = \{F_{V_t}^i|_{t=1}^T\}$, $i \in \{1,2,3\}$, T is number of frames. Given each audio frame $F_{A_t}^i$, the model needs to identify its corresponding visual frame $F_{V_t}^i$ from the entire visual sequence, and vice versa. In this sense, the A-V sequences can get well aligned to each other.

The within-layer contrastive loss is defined as:

$$\begin{aligned}
\mathcal{L}^{a2v}(F_A^i, F_V^i) &= -\sum_{t=1}^{T} \log \frac{\exp(\langle F_{A_t}^i, F_{V_t}^i \rangle / \tau)}{\sum_{n=1}^{T} \exp(\langle F_{A_t}^i, F_{V_n}^i \rangle / \tau)}, \\
\mathcal{L}^{v2a}(F_V^i, F_A^i) &= -\sum_{t=1}^{T} \log \frac{\exp(\langle F_{V_t}^i, F_{A_t}^i \rangle / \tau)}{\sum_{n=1}^{T} \exp(\langle F_{V_t}^i, F_{A_n}^i \rangle / \tau)}, \\
\mathcal{L}_{WL}^{i} &= \left[\mathcal{L}^{a2v}(F_A^i, F_V^i) + \mathcal{L}^{v2a}(F_V^i, F_A^i)\right] / 2,
\end{aligned} \tag{6}$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity, τ is temperature parameter. The two alignment directions (*i.e.*, $a2v$, $v2a$) are averaged to obtain the final WL contrastive loss.

Cross-Layer (CL) Contrastive Learning aims to align the A-V features across different GI model layers. As presented in Figure 2(a)(c), we select the j-th layer's output audio feature X_A^j and k-th layer's output visual feature X_V^k for alignment, where $j, k \in \{0,1,2,3\}$, $j \neq k$. Particularly, in this work we select $(j,k) \in \{(0,3),(3,0)\}$ to align the input and output A-V features of entire GI model, where more selections are discussed in ablation study (See Section 4.3).

Denote that $X_A^j = \{X_{A_t}^j|_{t=1}^T\}, X_V^k = \{X_{V_t}^k|_{t=1}^T\}$, where T is number of frames. First, we randomly sample T' A-V frame pairs from them for alignment, as a dropout to prevent over-fitting. Therefore, we can write the sampled frames as $\{(X_{A_t}^j, X_{V_t}^k)|t \in I\}$, where $I \subset \{1,2,...,T\}$, $|I| = T'$.

Then, we introduce a vector-quantization (VQ) module [Baevski *et al.*, 2019] to discretize the sampled audio-visual frames to a finite set of representations, which results in quantized targets to enable more effective contrastive learning, especially between different-layer features that usually locate in distant domains [Baevski *et al.*, 2020]:

$$Z_{A_t}^j = VQ(X_{A_t}^j), \quad Z_{V_t}^k = VQ(X_{V_t}^k), \quad t \in I, \tag{7}$$

Finally, we calculate cross-layer contrastive loss to align the audio/visual frames to the quantized visual/audio repre-

sentations respectively, similar to WL contrastive loss:

$$\mathcal{L}^{a2v}(X_A^j, Z_V^k) = -\sum_{t\in I} \log \frac{\exp(\langle X_{A_t}^j, Z_{V_t}^k\rangle/\tau)}{\sum_{n\in I_t}\exp(\langle X_{A_t}^j, Z_{V_n}^k\rangle/\tau)},$$

$$\mathcal{L}^{v2a}(X_V^k, Z_A^j) = -\sum_{t\in I} \log \frac{\exp(\langle X_{V_t}^k, Z_{A_t}^j\rangle/\tau)}{\sum_{n\in I_t}\exp(\langle X_{V_t}^k, Z_{A_n}^j\rangle/\tau)},$$

$$\mathcal{L}_{CL}^{j,k} = \left[\mathcal{L}^{a2v}(X_A^j, Z_V^k) + \mathcal{L}^{v2a}(X_V^k, Z_A^j)\right]/2, \tag{8}$$

where I_t contains the index t and another 100 randomly-selected indexes from I, for positive and negative samples respectively [Baevski *et al.*, 2020]. The two alignment directions are averaged to obtain the final CL contrastive loss.

3.4 Training Objective

We first calculate cross-entropy based sequence-to-sequence loss [Watanabe *et al.*, 2017] for speech recognition, as indicated by $\mathcal{L}_{ASR}$ in Figure 1(a). Then, we build the local alignment loss $\mathcal{L}_{LA}$ from WL and CL contrastive learning:

$$\mathcal{L}_{LA} = \sum_{i}^{M} \lambda_{WL}^{i} \cdot \mathcal{L}_{WL}^{i} + \sum_{(j,k)}^{N} \lambda_{CL}^{j,k} \cdot \mathcal{L}_{CL}^{j,k} \tag{9}$$

where $M = \{1, 2, 3\}$, $N = \{(0, 3), (3, 0)\}$, λ_{WL}^{i} and $\lambda_{CL}^{j,k}$ are weighting parameters for different training objectives.

We combine them to form the final training objective and train the entire GILA system in an end-to-end manner:

$$\mathcal{L}_{GILA} = \mathcal{L}_{ASR} + \mathcal{L}_{LA} \tag{10}$$

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on two large-scale publicly available datasets, LRS3 [Afouras *et al.*, 2018b] and LRS2 [Chung *et al.*, 2017]. LRS3 dataset collects 433 hours of transcribed English videos from TED and TEDx talks. LRS2 dataset contains 224 hours of video speech from BBC programs. More details can be found in [Shi *et al.*, 2022a].

Baselines. We employ AV-HuBERT[1] [Shi *et al.*, 2022a] as our baseline, but for fair comparison we discard the pre-training stage. To evaluate our GILA, we select some popular AVSR methods for comparison: TM-seq2seq, TM-CTC, Hyb-RNN, EG-seq2seq, RNN-T, LF-MMI TDNN, Hyb-Conformer, MoCo+wav2vec, AV-HuBERT (LARGE), u-HuBERT (LARGE), which are introduced in Section 2.

Implementation Details. For model configurations, our baseline follows AV-HuBERT LARGE [Shi *et al.*, 2022a] with 24 Transformer encoder layers and 9 decoder layers. For fair comparison, we build the GILA with 3 GI model layers, 12 Transformer encoder layers and 9 decoder layers. All other model configurations are same as AV-HuBERT LARGE. The number of parameters in our baseline and GILA are 476M and 465M respectively. We also use Conformer as our backbone, with the convolution kernel size of 31.

[1]https://github.com/facebookresearch/av_hubert

Method		Backbone	LM	WER(%) Clean	WER(%) Noisy
TM-seq2seq [2018a]		Transformer	✓	7.2	-
EG-seq2seq [2020]		RNN	-	6.8	-
RNN-T [2019]		RNN	-	4.5	-
Hyb-Conformer [2021]		Conformer	✓	2.3*	-
AV-HuBERT [2022a]		Transformer	-	1.4**	5.8**
u-HuBERT [2022]		Transformer	-	1.2**	-
GILA (ours)	Baseline	Transformer	-	3.75	17.22
	+ GI			3.29	15.06
	+ LA			2.88	13.35
	+ DA			2.61	11.14
	Baseline	Conformer	-	2.64	11.89
	+ GI			2.31	10.34
	+ LA			2.04	8.97
	+ DA			**1.96**	**7.03**

Table 1: WER (%) of GILA and prior works on LRS3 banchmark. "GI" denotes global interaction model, "LA" denotes local alignment approach, "DA" denotes data augmentation. "LM" denotes language model rescoring. * denotes using hybrid seq2seq/CTC loss for training, external LM rescoring for inference and extra data to pre-train the audio/visual front-ends. ** denotes using self-supervised pre-training with extra unlabeled data ($>$ 1,700 hours).

Method		Backbone	LM	WER(%) Clean	WER(%) Noisy
TM-seq2seq [2018a]		Transformer	✓	8.5	-
TM-CTC [2018a]		Transformer	✓	8.2	-
Hyb-RNN [2018]		RNN	✓	7.0	-
LF-MMI TDNN [2020]		TDNN	✓	5.9	-
Hyb-Conformer [2021]		Conformer	✓	3.7*	-
MoCo+wav2vec [2022]		Transformer	-	2.6**	-
GILA (ours)	Baseline	Transformer	-	5.79	25.52
	+ GI			4.98	21.91
	+ LA			4.31	18.84
	+ DA			4.02	15.70
	Baseline	Conformer	-	4.09	17.83
	+ GI			3.54	15.41
	+ LA			3.17	13.75
	+ DA			**3.10**	**11.24**

Table 2: WER (%) of our GILA and prior works on the LRS2 benchmark. * denotes the same as that in Table 1. ** denotes using self-supervised pre-trained audio/visual front-ends.

The system inputs are log filterbank features for audio stream and lip regions-of-interest (ROIs) for video stream. To sample A-V frame pairs in CL contrastive learning, we first sample starting indexes from (X_A^0, X_V^3) with probability of 0.4 and from (X_A^3, X_V^0) with 0.45 respectively, and then cut out 10 consecutive frames after each sampled index. To calculate contrastive loss, we use the same VQ module in wav2vec2.0 [Baevski *et al.*, 2020], and set the temperature parameter τ to 0.1. We further use data augmentation to improve noise robustness, where we add MUSAN noise [Snyder *et al.*, 2015] following prior work [Shi *et al.*, 2022b], and report WER results on both clean and noisy test sets. The weighting parameters $\lambda_{WL}^{i}(i \in \{1, 2, 3\})/\lambda_{CL}^{0,3}/\lambda_{CL}^{3,0}$ are set to 0.001/0.08/0.01 respectively. All hyper-parameters are tuned on validation set. Our training follows the finetuning configurations in [Shi *et al.*, 2022a] and takes $\sim$ 1.3 days on 4 V100-32GB GPUs, which is much more efficient than AV-

Method	Backbone	WER(%) Clean	WER(%) Noisy
Baseline	Transformer-LARGE	3.75	17.22
+ cross-attention		3.50	15.90
+ IR module		3.61	16.48
+ both (GI)		3.29	15.06
Baseline	Conformer-LARGE	2.64	11.89
+ cross-attention		2.45	10.94
+ IR module		2.53	11.41
+ both (GI)		**2.31**	**10.34**

Table 3: Effect of global interaction (GI) model and its two sub-modules on LRS3 benchmark. "+ cross-attention" denotes using cross-attention module separately, "+ IR module" denotes using iterative refinement module separately, where the self-attention and FFN modules in GI model are always maintained.

HuBERT pre-training ($\sim$ 15.6 days on 64 V100-GPUs). The details of data augmentation, model and training configurations follow previous work [Shi *et al.*, 2022b].

4.2 Main Results

Results on LRS3. Table 1 compares the performance of our proposed GILA with existing methods on LRS3 benchmark. Under clean test set, our best model outperforms the supervised learning SOTA by 14.8% relatively (2.3%→1.96%), while without the CTC training loss, external LM rescoring and extra A/V front-end pre-training that their method uses. Moreover, the proposed GILA has also achieved significant WER improvements over our baseline (3.75%→2.61%, 2.64%→1.96%). Specifically, its two main components, *i.e.*, GI model and LA method, both contribute a lot to the improvements, and the data augmentation also yields better results. We can also observe similar improvements on noisy test set. In addition, the Conformer backbone significantly outperforms Transformer (2.61%→1.96%).

Results on LRS2. Table 2 compares the performance of our GILA with existing AVSR methods on LRS2 benchmark. Under clean test set, our best model achieves 16.2% relative WER improvement over the supervised learning SOTA (3.7%→3.10%). Moreover, the GILA has also achieved significant improvements over our baseline (5.79%→4.02%, 4.09%→3.10%), where the GI model, LA method and data augmentation all yield positive contributions.

Therefore, our GILA has achieved new supervised learning SOTA on both LRS3 and LRS2 benchmarks, with up to 16.2% relative WER improvement over the best baseline. It also moves closer to the self-supervised learning SOTA (1.96% vs. 1.2%, 3.10% vs. 2.6%) while costs no unlabeled data and much less computing resources (See Section 4.1).

4.3 Ablation Study

Effect of Global Interaction Model. Table 3 summarizes the effect of proposed GI model and its two sub-modules, *i.e.*, cross-attention and IR modules. We first observe that using cross-attention to capture inter-modal correspondence can improve the WER results (3.75%→3.50%, 2.64%→2.45%). Further improvements can be achieved by adding IR module to jointly model the inter- and intra-modal correspondence

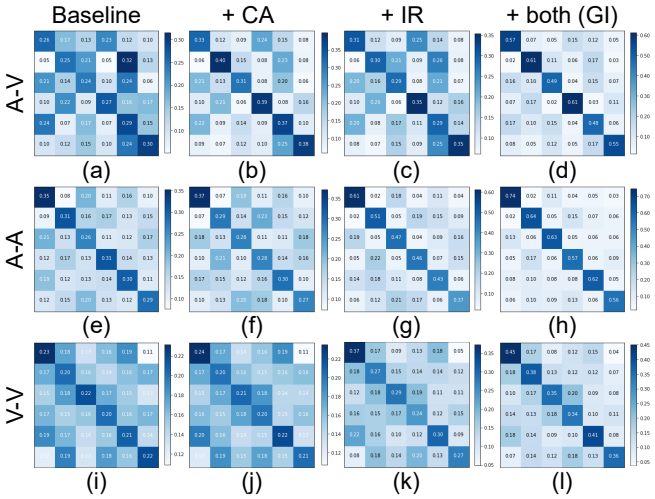

Figure 3: Cosine similarity matrix (after softmax) between audio-visual (row 1), audio-audio (row 2) and visual-visual (row 3) sequence embeddings in GI model. The column 1-4 denotes baseline, baseline + cross-attention, baseline + IR module, baseline + both (GI), respectively. In row 1, horizontal axis denotes visual sequences in a batch and vertical axis denotes the audio sequences, which are selected from LRS3 test set. Sequence embedding is obtained by temporal pooling on the output audio/visual sequences, *i.e.*, X_A^3, X_V^3.

(3.50%→3.29%, 2.45%→2.31%), where using it separately can also improve. Similar improvements can be observed on the noisy test set. Therefore, these results verify the effectiveness of our proposed GI model.

Visualizations of Inter- and Intra-Modal Correspondence. Figure 3 visualizes the captured inter- and intra-modal correspondence by our GI model, using similarity matrixes where the diagonal elements denote cosine similarity between true A-V, A-A or V-V pairs. We first observe chaotic mappings between A-V embeddings in baseline from Figure 3(a). After introducing cross-attention to interact A-V features, we can capture some inter-modal correspondence between true A-V pairs, *i.e.*, (b) vs. (a). However, it fails to capture the A/V intra-modal correspondence, *i.e.*, (f) vs. (e), (j) vs. (i). Thus, we further propose an iterative refinement module to jointly model the inter- and intra-modal correspondence, which improves significantly as indicated by the clearer diagonals in column 4. As a result, our GI model can effectively capture both inter- and intra-modal correspondence.

We further investigate the relationship between these two correspondences. When compared to baseline, using cross-attention can learn better inter-modal correspondence, *i.e.*, (b) vs. (a), while using it on top of IR module achieves significantly more improvements, *i.e.*, (d) vs. (c). Similar phenomenon can be observed on WER results in Table 3. It indicates that the proposed IR could be beneficial to cross-attention, where its captured intra-modal correspondence could help to model the inter-modal correspondence, thus results in better A-V complementary relationship.

Effect of Local Alignment Approach. Table 4 summarizes the effect of proposed LA method and its two components, *i.e.*, within-layer and cross-layer contrastive learning. We

Method	Backbone	WER(%) Clean	WER(%) Noisy
GI model	Transformer-LARGE	3.29	15.06
+ WL contrastive learning		3.03	13.92
+ CL contrastive learning		3.11	14.36
+ both (LA)		2.88	13.35
GI model	Conformer-LARGE	2.31	10.34
+ WL contrastive learning		2.13	9.53
+ CL contrastive learning		2.18	9.70
+ both (LA)		**2.04**	**8.97**

Table 4: Effect of local alignment (LA) approach and its two components on LRS3 benchmark.

WER(%)	X_V^0	X_V^1	X_V^2	X_V^3
X_A^0	-	2.12	2.11	**2.07**
X_A^1	2.12	-	2.12	2.09
X_A^2	2.09	2.10	-	2.11
X_A^3	**2.06**	2.08	2.10	-

Table 5: Effect of cross-layer contrastive learning. We select different A-V feature pairs (X_A^j, X_V^k) for cross-layer alignment. The baseline we use in this study is GI model with WL contrastive learning (2.13% WER in Table 4).

first introduce WL contrastive learning for audio-visual alignment within same GI model layer, which can improve the WER performance (3.29%→3.03%, 2.31%→2.13%). Further improvements can be achieved by adding CL contrastive learning to align the A-V features across different layers (3.03%→2.88%, 2.13%→2.04%), where using it separately can also improve. Similar improvements can be observed on noisy test set. Therefore, these results validate the effectiveness of our proposed LA method.

Effect of Cross-Layer Contrastive Learning. Table 5 further analyzes the effect of cross-layer contrastive learning, where we report WER results of alignment between different A-V feature pairs (X_A^j, X_V^k). We observe that the more layers our A-V alignment across (*i.e.*, larger $|j-k|$), the better performance we can achieve, where the best two results (2.07%, 2.06%) are achieved by aligning the input and output A-V features of entire GI model. After combining them, we can achieve even better WER result, as indicated in Table 4 (2.04%). The reason could be that, the higher-layer features contain semantic representations of larger granularity, or larger receptive field. Therefore, the A-V alignment across more layers also means across larger granularity gap, which could learn richer cross-modal contextual information and results in more informative A-V temporal consistency.

Visualizations of Audio-Visual Temporal Consistency. Figure 4 visualizes the A-V temporal consistency modeled by within-layer and cross-layer contrastive learning, using attention map where the diagonal elements indicate the attention weights between corresponding A-V frames. We first observe misalignment between A-V sequences in GI model, such as the one-to-many lip-audio mappings shown in Figure 4(a). Our proposed WL contrastive learning can help model the temporal consistency between A-V sequences, as indicated

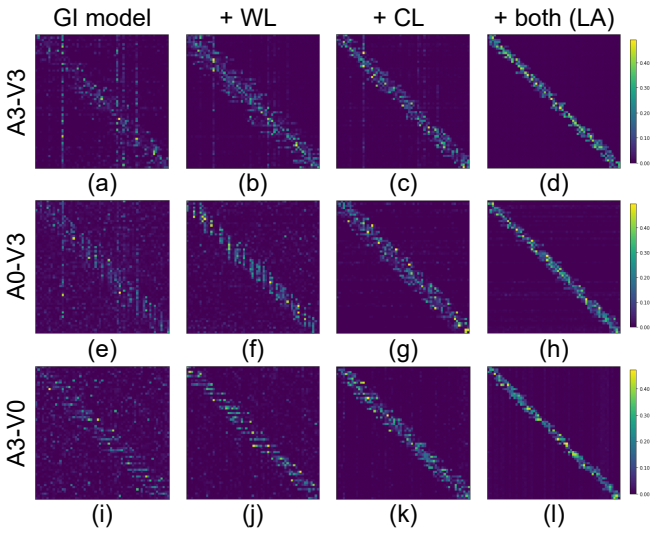

Figure 4: Attention weight map between different audio-visual sequences with LA method: row 1: (X_A^3, X_V^3), row 2: (X_A^0, X_V^3), row 3: (X_A^3, X_V^0). The column 1-4 denotes GI model, GI + WL contrastive learning, GI + CL contrastive learning, GI + both (LA), respectively. The x-axis denotes visual frames in an utterance and y-axis denotes the audio frames in utterance, which is selected from LRS3 test set.

by the clearer diagonal in (b). Similar improvements can be observed on cross-layer temporal consistency, *i.e.*, (f)/(j) vs. (e)/(i), while we also observe some vertical and horizontal stripes near the diagonal, which indicate the granularity gap between different-layer features.

Then in the proposed CL contrastive learning that consists of two alignment directions (See Equation 8), the low-layer features first learn rich A-V contextual correlations from the high-layer features that with large receptive field, which alleviates the granularity gap between them, *i.e.*, (g)/(k) vs. (e)/(i), (h)/(l) vs. (f)/(j). Meanwhile, the high-layer features can learn clearer A-V contextual mappings by aligned to the low-layer features that with small granularity, as indicated by the brighter diagonals in Figure 4 (column 3 vs. column 1, column 4 vs. column 2). As a result, the proposed cross-layer alignment can capture rich cross-modal contextual information to learn better A-V temporal consistency.

5 Conclusion

In this paper, we propose a cross-modal global interaction and local alignment (GILA) approach for audio-visual speech recognition, in order to capture the deep audio-visual correlations from both global and local perspectives. In particular, we first propose a global interaction model to capture the A-V complementary relationship on modality level. Furthermore, we design a cross-modal local alignment approach to model the A-V temporal consistency on frame level. Such a holistic view of cross-modal correlations enable better multimodal representations for AVSR. Experimental results on two public benchmarks demonstrate that our approach has achieved the state-of-the-art in supervised learning methods.

Acknowledgments

This research is supported by ST Engineering Mission Software & Services Pte. Ltd under collaboration programme (Research Collaboration No.: REQ0149132). The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

References

[Afouras *et al.*, 2018a] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[Afouras *et al.*, 2018b] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018.

[Ba *et al.*, 2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[Baevski *et al.*, 2019] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. In *International Conference on Learning Representations*, 2019.

[Baevski *et al.*, 2020] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[Biswas *et al.*, 2016] Astik Biswas, Prakash Kumar Sahu, and Mahesh Chandra. Multiple cameras audio visual speech recognition using active appearance model visual features in car environment. *International Journal of Speech Technology*, 19(1):159–171, 2016.

[Chopra *et al.*, 2005] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[Chung *et al.*, 2017] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3444–3453. IEEE, 2017.

[Glorot *et al.*, 2011] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

[Goncalves and Busso, 2022] Lucas Goncalves and Carlos Busso. Auxformer: Robust approach to audiovisual emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7357–7361. IEEE, 2022.

[Graves, 2012] Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012.

[Gu *et al.*, 2021] Chunbin Gu, Jiajun Bu, Zhen Zhang, Zhi Yu, Dongfang Ma, and Wei Wang. Image search with text feedback by deep hierarchical attention mutual information maximization. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4600–4609, 2021.

[Gulati *et al.*, 2020] Anmol Gulati, James Qin, Chiu Chung-Cheng, 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 *Interspeech*, pages 5036–5040, 2020.

[Hori *et al.*, 2017] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, pages 4193–4202, 2017.

[Hsu and Shi, 2022] Wei-Ning Hsu and Bowen Shi. u-hubert: Unified mixed-modal speech pretraining and zero-shot transfer to unlabeled modality. In *Advances in Neural Information Processing Systems*, 2022.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 448–456. PMLR, 2015.

[Kim *et al.*, 2022] Minsu Kim, Jeong Hun Yeo, and Yong Man Ro. Distinguishing homophenes using multi-head visual-audio memory for lip reading. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada*, volume 22, 2022.

[Koguchi *et al.*, 2018] Yuto Koguchi, Kazuya Oharada, Yuki Takagi, Yoshiki Sawada, Buntarou Shizuki, and Shin Takahashi. A mobile command input through vowel lip shape recognition. In *International Conference on Human-Computer Interaction*, pages 297–305. Springer, 2018.

[Korbar *et al.*, 2018] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, 31, 2018.

[Lee *et al.*, 2020] Jun-Tae Lee, Mihir Jain, Hyoungwoo Park, and Sungrack Yun. Cross-attentional audio-visual fusion for weakly-supervised action localization. In *International Conference on Learning Representations*, 2020.

[Li *et al.*, 2021] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[Lv *et al.*, 2021] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2554–2562, 2021.

[Ma *et al.*, 2021] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7613–7617. IEEE, 2021.

[Makino *et al.*, 2019] Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan. Recurrent neural network transducer for audio-visual speech recognition. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 905–912. IEEE, 2019.

[Mercea *et al.*, 2022] Otniel-Bogdan Mercea, Lukas Riesch, A Koepke, and Zeynep Akata. Audio-visual generalised zero-shot learning with cross-modal attention and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10553–10563, 2022.

[Morgado *et al.*, 2021] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021.

[Pan *et al.*, 2022] 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 *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4491–4503, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[Petridis *et al.*, 2018] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 513–520. IEEE, 2018.

[Raij *et al.*, 2000] Tommi Raij, Kimmo Uutela, and Riitta Hari. Audiovisual integration of letters in the human brain. *Neuron*, 28(2):617–625, 2000.

[Shi *et al.*, 2022a] 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.

[Shi *et al.*, 2022b] Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. Robust self-supervised audio-visual speech recognition. *arXiv preprint arXiv:2201.01763*, 2022.

[Snyder *et al.*, 2015] David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.

[Sumby and Pollack, 1954] William H Sumby and Irwin Pollack. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215, 1954.

[Tao and Busso, 2018] Fei Tao and Carlos Busso. Gating neural network for large vocabulary audiovisual speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(7):1290–1302, 2018.

[Tsai *et al.*, 2019] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2020] Zhong-Qiu Wang, Peidong Wang, and DeLiang Wang. Complex spectral mapping for single- and multi-channel speech enhancement and robust asr. *IEEE/ACM transactions on audio, speech, and language processing*, 28:1778–1787, 2020.

[Wang *et al.*, 2022] Daheng Wang, Tong Zhao, Wenhao Yu, Nitesh V Chawla, and Meng Jiang. Deep multimodal complementarity learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[Watanabe *et al.*, 2017] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253, 2017.

[Xu *et al.*, 2020] Bo Xu, Cheng Lu, Yandong Guo, and Jacob Wang. Discriminative multi-modality speech recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14433–14442, 2020.

[Yang *et al.*, 2022] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022.

[Yu *et al.*, 2020] 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 *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6984–6988. IEEE, 2020.