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

Yuchen Hu\textsuperscript{1}, Ruizhe Li\textsuperscript{2}, Chen Chen\textsuperscript{1}, Heqing Zou\textsuperscript{1}, Qiushi Zhu\textsuperscript{3} and Eng Siong Chng\textsuperscript{1}
\textsuperscript{1}Nanyang Technological University, Singapore
\textsuperscript{2}University of Aberdeen, UK
\textsuperscript{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 [Biswa 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 homophones [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 multi-modal 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. Hybrid-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. Hybrid-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.
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 \(F_V^i\):
\[
\begin{align*}
F_A^i &= LN(X_A^{i-1} + MHA(X_A^{i-1}, X_A^{i-1}, X_A^{i-1})), \\
F_V^i &= LN(X_V^{i-1} + MHA(X_V^{i-1}, X_V^{i-1}, X_V^{i-1})),
\end{align*}
\]
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:
\[
\begin{align*}
H_A^i &= LN(F_A^i + MHA(F_A^i, F_V^i, F_V^i)), \\
H_V^i &= LN(F_V^i + MHA(F_V^i, F_A^i, F_A^i)),
\end{align*}
\]
After that, we utilize position-wise feed-forward network (FFN) [Vaswani et al., 2017] to generate outputs:
\[
\begin{align*}
X_A^i &= LN(H_A^i + FFN(H_A^i)), \\
X_V^i &= LN(H_V^i + FFN(H_V^i)),
\end{align*}
\]
where FFN consists of two linear layers with a ReLU 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 \(R_V^i\):
\[
\begin{align*}
R_A^i &= \text{Conv}(\text{Attention}(X_{BN}^{i-1}, X_A^i, X_A^i)), \\
R_V^i &= \text{Conv}(\text{Attention}(X_{BN}^{i-1}, X_V^i, X_V^i)),
\end{align*}
\]
where “Conv” denotes a \(1 \times 1\) convolution layer followed by batch normalization (BN) [Ioffe and Szegedy, 2015] and parametric ReLU (PReLU) activation.
Within-Layer (WL) Contrastive Learning aims to align formative A-V temporal consistency on frame level. Based on that, we propose a novel layer contrastive learning to align the A-V features within the same GI model layer. Specifically, we first introduce within-model the temporal consistency between A-V frames, as presented in Figure 2. Specifically, we first introduce within-layer contrastive loss is defined as:

$$L_{WL}^{a2v}(F_A^i, F_V^i) = - \sum_{t=1}^{T} \log \frac{\exp(\langle F_A^i, F_V^i \rangle / \tau)}{\sum_{t=1}^{T} \exp(\langle F_A^i, F_V^i \rangle / \tau)},$$

$$L_{WL}^{v2a}(F_V^i, F_A^i) = - \sum_{t=1}^{T} \log \frac{\exp(\langle F_V^i, F_A^i \rangle / \tau)}{\sum_{t=1}^{T} \exp(\langle F_V^i, F_A^i \rangle / \tau)},$$

$$L_{VL}^{WL} = \frac{L_{WL}^{a2v}(F_A^i, F_V^i) + L_{WL}^{v2a}(F_V^i, F_A^i)}{2},$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity, $\tau$ 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 $X_A^j = \{X_A^{j}_{i=1} [t]\}$, $X_V^k = \{X_V^{k}_{i=1} [t]\}$, 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 $\{(X_A^{j}_{i}, X_V^{k}_{i})|t \in I\}$, where $I \subset \{1, 2, \ldots, T\}$, $|I| = T'$.

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]:

$$Z_A^{j} = VQ(X_A^{j}), \quad Z_V^{k} = VQ(X_V^{k}), \quad t \in I,$$

Finally, we calculate cross-layer contrastive loss to align the audio/visual frames to the quantized visual/audio representations.
sentations respectively, similar to WL contrastive loss:
\[
\mathcal{L}^{2v}(X^r_A, Z^v_V) = - \sum_{t \in I} \log \frac{\exp(\langle X^r_A, Z^v_V \rangle / \tau)}{\sum_{n \in I_t} \exp(\langle X^r_A, Z^v_V \rangle / \tau)},
\]
\[
\mathcal{L}^{v2a}(X^v_V, Z^v_A) = - \sum_{t \in I} \log \frac{\exp(\langle X^v_V, Z^v_A \rangle / \tau)}{\sum_{n \in I_t} \exp(\langle X^v_V, Z^v_A \rangle / \tau)},
\]
\[
\mathcal{L}^{i,k}_{CL} = \left[ \mathcal{L}^{2v}(X^r_A, Z^v_V) + \mathcal{L}^{v2a}(X^v_V, Z^v_A) \right] / 2,
\]
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 \(L_{ASR}\) in Figure 1(a). Then, we build the local alignment loss \(L_{LA}\) from WL and CL contrastive learning:
\[
L_{LA} = \sum_{i}^{M} \lambda^{i}_{WL} \cdot \mathcal{L}^{WL} + \sum_{(j,k)}^{N} \lambda^{i,k}_{CL} \cdot \mathcal{L}^{i,k}_{CL}
\]
where \(M = \{1, 2, 3\}, N = \{(0, 3), (3, 0)\}\), \(\lambda^{i}_{WL}\) and \(\lambda^{i,k}_{CL}\) 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:
\[
L_{GILA} = L_{ASR} + L_{LA}
\]

### 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 [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 ASVR 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.

---

1https://github.com/facebookresearch/av_hubert
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%→3.11%). 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 (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 matrices 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...
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 \((X^j_A, X^k_V)\). We observe that the more layers our A-V alignment across \((i.e., \text{larger } j - k)\), the better performance we can achieve, where the last 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 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


