Abstract

Audio-visual speech recognition (AVSR) attracts a surge of research interest recently by leveraging multimodal signals to understand human speech. Mainstream approaches addressing this task have developed sophisticated architectures and techniques for multi-modality fusion and representation learning. However, the natural heterogeneity of different modalities causes distribution gap between their representations, making it challenging to fuse them. In this paper, we aim to learn the shared representations across modalities to bridge their gap. Different from existing similar methods on other multimodal tasks like sentiment analysis, we focus on the temporal contextual dependencies considering the sequence-to-sequence task setting of AVSR. In particular, we propose an adversarial network to refine frame-level modality-invariant representations (MIR-GAN), which captures the commonality across modalities to ease the subsequent multimodal fusion process. Extensive experiments on public benchmarks LRS3 and LRS2 show that our approach outperforms the state-of-the-arts.

1 Introduction

Human perception of the world intrinsically comprises multiple modalities, including vision, audio, text, etc. (McGurk and MacDonald, 1976; Bartušaitis et al., 2018). Audio-visual speech recognition (AVSR) leverages both audio and visual modalities to understand human speech, improving the noise-robustness of audio-only speech recognition (Chen et al., 2022a,b, 2023a,b; Hu et al., 2022b,a, 2023a,c,b; Zhu et al., 2023a,b) with noise-invariant lip movement information (Sumby and Pollack, 1954). Thanks to recent advances of deep learning techniques, AVSR research has gained a remarkable progress (Afouras et al., 2018a; Ma et al., 2021; Shi et al., 2022b).

Currently, the mainstream AVSR approaches are centered around developing sophisticated architectures and techniques for multi-modality fusion, including simple feature concatenation (Makino et al., 2019; Ma et al., 2021; Pan et al., 2022; Chen et al., 2022c; Zhu et al., 2023c), recurrent neural network (Petridis et al., 2018; Xu et al., 2020) and cross-modal attention (Afouras et al., 2018a; Lee et al., 2020; Hu et al., 2023d). Despite the advances, these approaches are often challenged by the representation gap persisting between naturally heterogeneous modalities (Hazarika et al., 2020). Recently in some other multimodal tasks like sentiment analysis (Hazarika et al., 2020; Yu et al., 2021; Yao and Mihalcea, 2022) and cross-modal retrieval (Xiong et al., 2020), there have been research works proposing to learn two distinct representations to benefit multimodal learning. The first representation is modality-invariant, where multiple modalities of a same utterance are mapped to a shared space, indicating the homogeneous semantic meaning from the speaker. In addition, they also learn modality-specific representations that are private to each modality. Given an utterance, each modality contains some unique features with respect to speaker-sensitive information (Tsiros, 2013). Combining these two representations provides a holistic view of multimodal data for downstream tasks (Yang et al., 2022). However, these meth-
ods focus on utterance-level representations that could be easily mapped to either shared or individual modality space using similarity cost functions, which does not apply to AVSR task that requires sequence-to-sequence mapping with temporal contextual dependencies (Petridis et al., 2018).

Motivated by above observations, we propose an adversarial network to refine frame-level modality-invariant representations (MIR-GAN) for capturing the commonality across modalities, which bridges their heterogeneous gap to ease the subsequent multimodal fusion. In particular, we first design a MIR generator to learn modality-invariant representations over the shared audio-visual modality space. Meanwhile, a modality discriminator is proposed to strengthen its modality agnosticism via adversarial learning. Moreover, to further enrich its contextual semantic information, we propose a mutual information maximization strategy to align the refined representations to both audio and visual modality sequences. Finally, both modality-invariant and -specific representations are fused for downstream speech recognition. Empirical results demonstrate the effectiveness of our approach. In summary, our main contributions are:

• We present MIR-GAN, an AVSR approach to refine frame-level modality-invariant representations, which captures the commonality across modalities and thus bridges their heterogeneous gap to ease multimodal fusion.

• We first learn modality-invariant representations with a MIR generator, followed by another modality discriminator to strengthen its modality agnosticism via adversarial learning. Furthermore, we propose a mutual information maximization strategy to enrich its contextual semantic information. Finally, both modality-invariant and -specific representations are fused for downstream recognition.

• Our proposed MIR-GAN outperforms the state-of-the-arts on LRS3 and LRS2 benchmarks. Extensive experiments also show its superiority on ASR and VSR tasks.

2 Related Work

Audio-Visual Speech Recognition. Current mainstream AVSR methods focus on sophisticated architectures and techniques for audio-visual modality fusion. Prior methods like RNN-T (Makino et al., 2019), Hyb-Conformer (Ma et al., 2021) and MoCo+wav2vec (Pan et al., 2022) employ simple feature concatenation for multimodal fusion, other works including Hyb-RNN (Petridis et al., 2018) and EG-seq2seq (Xu et al., 2020) leverage recurrent neural network for audio-visual fusion. In addition, cross-modal attention has also become popular recently for multimodal interaction and fusion in AVSR tasks, such as TM-seq2seq (Afouras et al., 2018a), DCM (Lee et al., 2020) and MMST (Song et al., 2022). Despite the effectiveness, these fusion techniques are often challenged by the representation gap between naturally heterogeneous modalities. Recently, multimodal self-supervised learning has been popular for capturing unified cross-modal representations, like AV-HuBERT (Shi et al., 2022a) and u-HuBERT (Hsu and Shi, 2022), which achieve the state-of-the-art but require abundant unlabeled data and computing resources. In this work, we propose a supervised learning scheme to efficiently refine modality-invariant representations for bridging the heterogeneous modality gap.

Modality-Invariant and -Specific Representations. Recent studies in many multimodal tasks suggest that the model benefits from both shared and individual modality representations, including multimodal sentiment analysis (Hazarika et al., 2020; Yu et al., 2021; Yang et al., 2022), person re-identification (Wei et al., 2021; Huang et al., 2022), cross-modal retrieval (Zeng et al., 2022) and image-sentence matching (Liu et al., 2019), etc. MISA (Hazarika et al., 2020) maps the multimodal features into two spaces as modality-invariant and -specific representations, and then fuses them for downstream classification. MCLNet (Hao et al., 2021) learns modality-invariant representations by minimizing inter-modal discrepancy and maximizing cross-modal similarity. VI-REID (Feng et al., 2019) builds an individual network for each modality, with a shared identity loss to learn modality-invariant representations. However, these methods map utterance-level representations to modality-invariant or -specific spaces via similarity cost functions, while AVSR is sequence-to-sequence task that requires contextual semantic information. To this end, we propose an adversarial network with mutual information maximization to refine frame-level modality-invariant representations that subjects to temporal contextual dependencies.

Adversarial Network. The concept of adversarial network starts from GAN (Goodfellow et al., 2014),
which has attracted a surge of research interests due to its strong ability of generating high-quality novel samples according to existing data. The best-known applications include image-to-image translation (Isola et al., 2017) and image synthesis (Denton et al., 2015; Radford et al., 2015). Recently, GAN is further applied to multimodal tasks such as text-to-image synthesis (Reed et al., 2016; Tan et al., 2020), video captioning (Yang et al., 2018; Bai et al., 2021) and cross-modal retrieval (Qian et al., 2021). In this work, we leverage the strong distinguishing ability of adversarial network to strengthen the modality agnosticism of the learned modality-invariant representations.

3 Methodology

3.1 Overview

The overall architecture of our proposed MIR-GAN is illustrated in Fig. 2. First, we have two front-end modules\(^2\) to process the input streams, which generate two modality sequences, i.e., \(f_v, f_a \in \mathbb{R}^{T \times D}\), where \(T\) is number of frames and \(D\) is embedding size. These two sequences are then fed by visual and audio encoders respectively to generate modality-specific representations, i.e., \(f_{v}^{spe}, f_{a}^{spe} \in \mathbb{R}^{T \times D}\). Based on that, we propose a MIR generator to learn modality-invariant representations by extracting the shared information of two modalities, i.e., \(f_{inv}^{va} \in \mathbb{R}^{T \times D}\). Meanwhile, we design a modality discriminator to strengthen its modality agnosticism via adversarial learning. In addition, to further enrich its contextual semantic information, we propose a mutual information maximization (MIM) strategy to align the refined representations to both audio and visual modality sequences. Finally, both modality-invariant and -specific representations are fused for downstream speech recognition.

3.2 Visual & Audio Encoders

As illustrated in Fig. 2 (d), we introduce a pair of visual and audio encoders to learn modality-specific representations. Following Transformer (Vaswani et al., 2017) architecture, they first employ self-attention modules to capture the contextual dependencies within each modality, followed by cross-attention modules for interaction between two modalities, which can initially narrow their gap to benefit the subsequent modality-invariant representation learning. Finally, there are feed-forward networks to generate the modality-specific outputs.

3.3 MIR-GAN

With learned modality-specific representations, we propose MIR-GAN to refine frame-level modality-invariant representations. First, we design a MIR

---

\(^2\)Details are presented in Appendix A.3.
generator to extract the shared information of two modalities, which generates a modality-invariant representation $f_{va}^{inv} \in \mathbb{R}^{T \times D}$. Meanwhile, we design a modality discriminator to strengthen its modality agnosticism via adversarial learning.

### 3.3.1 MIR Generator

Fig. 2 (b) details the architecture of proposed MIR generator $G$, where we design a hybrid-modal attention (HMA) module to extract out the part of information in each modality-specific representation that is related to both modalities:

$$s_m = HMA(f_{va}^{spe}, f_{va}), \quad m \in \{v, a\},$$  (1)

where the subscript $m$ denotes modality. The resulted features are then added to input sequence $f_{va}$ to form the final modality-invariant representation:

$$f_{va}^{inv} = \text{Norm}(f_{va} + \sum_{m \in \{v, a\}} \text{Conv}(s_m)), \quad (2)$$

where the “Norm” denotes layer normalization (Ba et al., 2016), “Conv” denotes 1x1 convolution followed by PReLU activation (He et al., 2015).

**Hybrid-Modal Attention (HMA)** first involves a cross-attention sub-module to extract the information in each modality-specific representation that is related to both modalities, with the query input $f_{va}$ comprising both visual and audio sequence information, as shown in Fig. 3:

$$f_m^{share} = \text{Cross-Attention}(f_{va}, f_{va}^{spe}, f_{va}^{spe}), \quad (3)$$

To further make the extracted feature invariant to modalities, we design a parallel convolutional network to learn a mask for filtering out the modality-specific information:

$$s_m = f_m^{share} \otimes \sigma(\text{Conv}(f_m^{spe} || f_{va})), \quad (4)$$

where “Conv” denotes 1x1 convolutional layer, $\sigma$ denotes feature concatenation, $\otimes$ denotes Sigmoid activation, $\sigma$ denotes element-wise multiplication.

As a result, the output representation $s_m$ from HMA involves information regarding both visual and audio modalities, making the final output $f_{va}^{inv}$ (in Eq. 2) invariant to modalities.

### 3.3.2 Modality Discriminator

With the generated modality-invariant representation, we further design a modality discriminator $D$ to strengthen its modality agnosticism via adversarial learning. As shown in Fig. 2 (c), the discriminator consists of two linear layers followed by Sigmoid activation to predict a scalar between 0 and 1 for each frame, indicating which modality it belongs to (i.e., 0 for visual and 1 for audio):

$$D(f) \in \mathbb{R}^{T \times 1}, \quad f \in \{f_v^{spe}, f_a^{spe}, f_{va}^{inv}\}, \quad (5)$$

Therefore, for frames in modality-specific representations $f_v^{spe}$ and $f_a^{spe}$, we hope the discriminator can correctly classify the modality type, i.e., 0 or 1. In contrast, in order to strengthen the modality agnosticism of refined representation $f_{va}^{inv}$, we hope it can confuse the discriminator with the output around 0.5, i.e., a medium between two modalities.

With above designs of generator and discriminator, the adversarial training objective of MIR-GAN can be mathematically formulated as:

$$\mathcal{L}_{GAN} = \mathcal{L}_D + \mathcal{L}_G$$

$$= \mathbb{E}_f[\log D(f_v^{spe}) + \log(1 - D(f_a^{spe}))]$$

$$+ \mathbb{E}_f[-\log D(f_{va}^{inv}) - \log(1 - D(f_{va}^{inv}))], \quad (6)$$

where $f_{va}^{inv} = G(f_v^{spe}, f_a^{spe}, f_{va})$. $\mathbb{E}$ denotes the expectation over all the temporal frames in current data batch. Details of the corresponding optimization strategy are illustrated in Alg. 1.

### 3.4 Mutual Information Maximization

The MIR-GAN successfully refines the modality-invariant representation by focusing on the modality commonality and agnosticism, while the original semantic information may not be preserved. To this end, we further design a mutual information maximization (MIM) strategy via contrastive learning to enrich the contextual semantic information in refined modality-invariant representation.

In particular, we formulate a contrastive loss function to maximize the mutual information between modality-invariant representation $f_{va}^{inv}$ and the modality-specific representations $f_v^{spe}, f_a^{spe}$.
where \( \langle \cdot, \cdot \rangle \) denotes cosine similarity, \( \tau \) is temperature parameter. The subscripts \( i \) and \( j \) denote frame index, where \( f^{\text{inv}}_i / f^{\text{spe}}_i / f^{\text{spe}}_a \in \mathbb{R}^{T \times D} \).

The constructed positive and negative samples are distinguished by frame index. As same frame of different representations express similar semantic meanings, we assign them as positive samples to strengthen consistency, while the mismatched frames are pulled apart from each other. As a result, the MIM strategy can enrich the semantic information in final modality-invariant representation.

### 3.5 Optimization

The optimization strategy of MIR-GAN is detailed in Alg. 1. After the forward-propagation process, we calculate \( \mathcal{L}_{\text{GAN}} \) and \( \mathcal{L}_{\text{MIM}} \) according to Eq. 6 and Eq. 7. Meanwhile, the downstream speech recognition loss \( \mathcal{L}_{\text{rec}} \) is calculated as the cross-entropy between recognized text and the ground-truth transcription. The final training objective of MIR-GAN can therefore be written as:

\[
\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{GAN}} \cdot \mathcal{L}_{\text{GAN}} + \lambda_{\text{MIM}} \cdot \mathcal{L}_{\text{MIM}},
\]

where \( \lambda_{\text{GAN}}, \lambda_{\text{MIM}} \) are weighting parameters to balance different training objectives.

Inspired by GAN training strategy (Goodfellow et al., 2014), we split the back-propagation process into two steps. First, we maximize \( \mathcal{L}_{\text{GAN}} \) to update the discriminator, where the generator is detached from optimization. According to Eq. 6, maximizing the first term of \( \mathcal{L}_{\text{GAN}} \) (i.e., \( \mathcal{L}_D \)) trains the discriminator to correctly classify the two modalities, while increasing the second term amounts to informing discriminator that \( f^{\text{inv}}_v \) is modality-specific and can either be visual or audio\(^3\) (this is opposite to what we desire as modality-agnostic). Second, we freeze discriminator and update the rest network, where minimizing \( \mathcal{L}_G \) pushes the discrimination output of \( f^{\text{inv}}_v \) to 0.5,\(^3\) which is a medium between visual and audio modalities, i.e., modality-agnostic.

---

**Algorithm 1 MIR-GAN Optimization.**

**Require:** Training data \( D \) that contains visual-audio pairs \((x_v, x_a)\) and the text transcription \( y \). The MIR-GAN network \( \theta \) consists of front-ends \( \theta_{\text{fe}} \) and \( \theta_{\text{fe}} \), encoders \( \theta_{\text{enc}} \), MIR generator \( \theta_G \), modality discriminator \( \theta_D \), and downstream speech recognition model \( \theta_{\text{rec}} \).

1: Randomly initialize the entire system \( \theta \).
2: if select self-supervised setting then
3: Load the pre-trained AV-HuBERT for speech recognition model \( \theta_{\text{rec}} \) and front-ends \( \theta_{\text{fe}} \).
4: end if
5: while not converged do
6: for \((x_v, x_a) \in D\) do
7: \textbf{FORWARD-PROPAGATION:}\n8: \( f_v = \theta_{\text{fe}}(x_v), f_a = \theta_{\text{fe}}(x_a) \) \triangleright \text{front-ends} \n9: f^{\text{spe}}_v, f^{\text{spe}}_a = \theta_{\text{enc}}(f_v, f_a) \triangleright \text{encoders} \n10: f^{\text{inv}}_v = f_v \parallel f_a \triangleright \text{Generator} \n11: f^{\text{inv}}_a = \theta_{\text{G}}(f^{\text{spe}}_v, f^{\text{spe}}_a, f^{\text{inv}}_a) \triangleright \text{Generator} \n12: \hat{y} = \theta_{\text{rec}}(f^{\text{spe}}_v \parallel f^{\text{spe}}_a \parallel f^{\text{inv}}_v) \triangleright \text{recognition} \n13: \textbf{TRAINING OBJECTIVES:}\n14: \mathcal{L}_{\text{GAN}} (\mathcal{L}_D \text{ and } \mathcal{L}_G) \text{ in Eq. 6} \triangleright \text{Discriminator} \n15: \mathcal{L}_{\text{MIM}} \text{ in Eq. 7} \triangleright \text{MIM maximization} \n16: \mathcal{L}_{\text{rec}} = \text{CrossEntropy}(\hat{y}, y) \triangleright \text{adversarial training} \n17: \textbf{BACK-PROPAGATION:}\n18: \text{UPDATE DISCRIMINATOR:} \triangleright \text{unfreeze } \theta_D \n19: \arg\max_{\theta_D, \theta_G} \mathcal{L}_{\text{GAN}} \n20: \text{UPDATE THE REST NETWORK:} \triangleright \text{freeze } \theta_D \n21: \arg\min_{\theta_{\text{fe}}, \theta_{\text{enc}}} \mathcal{L}_{\text{rec}} + \lambda_{\text{GAN}} \cdot \mathcal{L}_{\text{GAN}} + \lambda_{\text{MIM}} \cdot \mathcal{L}_{\text{MIM}} \n22: \text{end for} \n23: \text{end while} \n
In addition, \( \mathcal{L}_{\text{rec}} \) optimizes the downstream speech recognition model and \( \mathcal{L}_{\text{MIM}} \) implements the MIM strategy. The entire system is trained in an end-to-end manner with well-tuned weighting parameters.

### 4 Experiments

#### 4.1 Experimental Setup

**Datasets.** We conduct experiments on two large-scale public benchmarks, LRS3 (Afouras et al., 2018b) and LRS2 (Chung et al., 2017). LRS3 dataset collects 433 hours of transcribed English videos in TED and TEDx talks from over 5000 speakers, which is the largest publicly available labeled audio-visual speech recognition dataset. LRS2 dataset contains 224 hours of video speech, with a total of 144K clips from BBC programs.

**Model Configurations.** We first build a base model with only front-ends and downstream speech recognition module, which follows Transformer architecture with 24 encoder layers and 9 decoder layers. Based on that, we build the MIR-GAN with \( N_E = 3 \) visual & audio encoder layers and \( N_G = 3 \) MIR generator layers. To maintain similar model size, we only use 12 encoder layers and 9 decoder layers in the recognition model. The number

---

\( ^3 \)Function \( \log(x) + \log(1-x) \) reaches maximum at \( x = 0.5 \), and the minimum is obtained around \( x = 0 \) and \( x = 1 \).
Table 1: WER (%) of our MIR-GAN and prior works on LRS3 benchmark. “S2S” denotes sequence-to-sequence loss (Watanabe et al., 2017), “CTC” denotes CTC loss (Graves et al., 2006), “DataAug” denotes noise augmentation, “LM” denotes language model rescoring. The noisy test set is synthesized using MUSAN noise (Snyder et al., 2015).

of parameters in our base model and MIR-GAN are 476M and 469M respectively. We also use Conformer (Gulati et al., 2020) as our backbone. In addition, we implement a self-supervised setting by loading pre-trained AV-HuBERT\(^4\). Following prior work (Shi et al., 2022b), we employ data augmentation and noisy test set based on MUSAN noise (Snyder et al., 2015). More detailed settings are presented in Appendix A.3 - A.5.

Baselines. To evaluate our proposed MIR-GAN, we select some popular AVSR methods for comparison, which can be roughly divided into two groups. The first is supervised learning method, including TM-seq2seq/CTC (Afouras et al., 2018a), RNN-T (Makino et al., 2019), EG-seq2seq (Xu et al., 2020) and Hyb-Conformer (Ma et al., 2021). Another one is the recently popular self-supervised learning method such as MoCo+wav2vec (Pan et al., 2022), AV-HuBERT (Shi et al., 2022b) and u-HuBERT (Hsu and Shi, 2022).

4.2 Main Results

We conduct experiments on two public datasets under supervised and self-supervised settings, depending on whether use the AV-HuBERT pre-trained model. Results show that our proposed MIR-GAN achieves the state-of-the-art under both settings.

LRS3 Benchmark. Table 1 presents the AVSR

\(^4\)https://github.com/facebookresearch/av_hubert

\[\text{Table 1: WER (%) of our MIR-GAN and prior works on LRS3 benchmark. “S2S” denotes sequence-to-sequence loss (Watanabe et al., 2017), “CTC” denotes CTC loss (Graves et al., 2006), “DataAug” denotes noise augmentation, “LM” denotes language model rescoring. The noisy test set is synthesized using MUSAN noise (Snyder et al., 2015).}\]

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Criterion</th>
<th>Unlabeled data (hrs)</th>
<th>Labeled data (hrs)</th>
<th>DataAug</th>
<th>LM</th>
<th>WER(%) Clean</th>
<th>WER(%) Noisy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TM-seq2seq (2018a)</td>
<td>Transformer</td>
<td>S2S</td>
<td>- 1,519</td>
<td>✓</td>
<td>✓</td>
<td>7.2</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>EG-seq2seq (2020)</td>
<td>RNN</td>
<td>S2S</td>
<td>- 590</td>
<td>✓</td>
<td>-</td>
<td>6.8</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>RNN-T (2019)</td>
<td>RNN</td>
<td>RNN-T</td>
<td>- 31,000</td>
<td>-</td>
<td>-</td>
<td>4.5</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Hyb-Conformer (2021)</td>
<td>Conformer</td>
<td>S2S + CTC</td>
<td>- 590</td>
<td>✓</td>
<td>✓</td>
<td>2.3</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Self-Supervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AV-HuBERT (2022b)</td>
<td>Transformer</td>
<td>S2S</td>
<td>1,759</td>
<td>433</td>
<td>✓</td>
<td>- 1.4</td>
<td>5.8</td>
<td></td>
</tr>
<tr>
<td>u-HuBERT (2022)</td>
<td>Transformer</td>
<td>S2S</td>
<td>2,211</td>
<td>433</td>
<td>✓</td>
<td>- 1.2</td>
<td>5.6</td>
<td></td>
</tr>
</tbody>
</table>

| Proposed (Supervised)   |          |           |                      |                   |         |    |              |              |
| Ours                    | Base model Transformer | S2S   | - 433               | ✓                 | -       | 3.5 | 14.8         |              |
|                         | MIR-GAN   |          |                      |                   |         |    |              |              |
| Proposed (Self-Supervised) |        |           |                      |                   |         |    |              |              |
| Ours                    | Base model Transformer | S2S   | 1,759               | 433               | ✓       | - 1.4| 5.8          |              |
|                         | MIR-GAN   |          |                      |                   |         |    |              |              |

Table 2: WER (%) of our MIR-GAN and prior works on the LRS2 benchmark. Detailed configurations are further presented in Table 6.

\[\text{Table 2: WER (%) of our MIR-GAN and prior works on the LRS2 benchmark. Detailed configurations are further presented in Table 6.}\]

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>WER(%) Clean</th>
<th>WER(%) Noisy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TM-seq2seq (2018a)</td>
<td>Transformer</td>
<td>8.5</td>
<td>-</td>
</tr>
<tr>
<td>TM-CTC (2018a)</td>
<td>Transformer</td>
<td>8.2</td>
<td>-</td>
</tr>
<tr>
<td>Hyb-RNN (2018)</td>
<td>RNN</td>
<td>7.0</td>
<td>-</td>
</tr>
<tr>
<td>LF-MMI TDNN (2020)</td>
<td>TDNN</td>
<td>5.9</td>
<td>-</td>
</tr>
<tr>
<td>Hyb-Conformer (2021)</td>
<td>Conformer</td>
<td>3.7</td>
<td>-</td>
</tr>
<tr>
<td>Self-Supervised</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MoCo+wav2vec (2022)</td>
<td>Transformer</td>
<td>2.6</td>
<td>-</td>
</tr>
</tbody>
</table>

| Proposed (Supervised)   |          |              |              |
| Ours                    | Base model Transformer | 5.4          | 21.2         |
|                         | MIR-GAN   | 4.5          | 16.7         |
| Proposed (Self-Supervised) |        |              |              |
| Ours                    | Base model Transformer | 3.9          | 13.8         |
|                         | MIR-GAN   | 3.2          | 11.9         |
| Ours                    | Base model Transformer | 2.3          | 7.3          |
|                         | MIR-GAN   | 2.2          | 7.0          |

performance of our proposed MIR-GAN and prior methods on LRS3 benchmark. Under supervised setting, our MIR-GAN achieves significant improvement over the base model in both clean and noisy testing conditions, and the best performance achieves new state-of-the-art (2.1% vs. 2.3%) while without using the language model rescoring. In addition, the Conformer backbone consistently outperforms Transformer (2.1% vs. 2.8%, 8.5% vs. 11.7%). Under self-supervised setting, MIR-
Table 3: Ablation study on LRS3 and LRS2 benchmarks. Results are reported on six configurations in the format "[Backbone]-[Setting]-[Test set]", where “TF”/“CF” denote Transformer/Conformer backbone, “Sup”/“SelfSup” denote supervised/self-supervised setting, “3”/“2” denote LRS3/LRS2 test set.

GAN also improves the performance of base model, which surpasses or matches previous state-of-the-art (1.2% vs. 1.2%, 5.6% vs. 5.8%) while using less unlabeled data for pre-training.

LRS2 Benchmark. Table 2 compares the AVSR results of MIR-GAN and baselines on LRS2 benchmark. We can observe that the proposed MIR-GAN outperforms previous state-of-the-art by a large margin under both supervised and self-supervised settings (3.2% vs. 3.7%, 2.2% vs. 2.6%). In addition, we also observe promising gains of performance in noisy testing conditions.

As a result, our proposed MIR-GAN achieves new state-of-the-art under both supervised and self-supervised settings on two public benchmarks, which demonstrates its superiority on AVSR task.

4.3 Ablation Study

Table 3 presents the ablation study of each component in MIR-GAN. There are three parts of ablation that are independent with each other, i.e., each study is conducted where other two components are kept same as the full MIR-GAN.

Importance of Representations. We first investigate the importance of modality-invariant and -specific representations by discarding each of them. When removing the refined modality-invariant representations from multi-modality fusion, the downstream speech recognition performance degrades a lot under all configurations, which verifies its significance of bridging the modality gap. Similarly, we observe that the modality-specific representations also plays an important role in AVSR.

Importance of Modules. In this part, we study the role of each module in the proposed MIR-GAN.

Importance of Strategies. With the adversarial training strategy illustrated in Alg. 1, the proposed modality discriminator effectively strengthens the modality agnosticism of the refined representations from generator. To verify its effectiveness, we remove the adversarial training strategy from MIR-GAN, which results in similar performance degradation to the previous case without modality discriminator. Therefore, it demonstrates the key role of this strategy in learning modality-invariant representations, where further visualization is shown in Fig. 5. Meanwhile, we design a MIM strategy to enrich the contextual semantic information in the
refined modality-invariant representations, and similar performance drops can be observed in absence of such strategy. Furthermore, we visualize the attention maps in Fig. 4 to show its effectiveness. The clear diagonals in (c) and (d) indicate the strong ability of MIM strategy to align modality-invariant and -specific representations, which enriches the contextual semantic information in the former.

**Visualizations of Modality-Invariant and -Specific Representations.** Fig. 5 presents the t-SNE visualization of modality-invariant and -specific representations to illustrate the principle of MIR-GAN. First, we observe from (a) base model that the two modality-specific representations are distinctly separated, indicating the heterogeneous gap between different modalities (Hazarika et al., 2020). With the proposed MIR-GAN (no modality discriminator), the two modalities are pushed closer by the interaction between encoders, and the refined modality-invariant representations serve as a medium between them. However, these refined representations are still entangled with audio modality-specific representations, making them less modality-invariant. Finally, the proposed discriminator effectively strengthens their modality agnosticism via adversarial learning, which are dispersed between two modalities to capture their commonality and thus bridge the heterogeneous modality gap. As a result, the subsequent multi-modality fusion process would be eased and generate better features for downstream recognition.

**Comparison with Utterance-Level Approaches**

As illustrated in §2, prior works have investigated utterance-level modality-invariant and -specific representations with similarity cost functions, including MISA (Hazarika et al., 2020), MCLNet (Hao et al., 2021) and VI-REID (Feng et al., 2019). We implement them in our framework as comparison to our proposed MIR-GAN, where we employ their designed similarity cost functions on frame-level representations. As illustrated in Table 4, these utterance-level approaches can also improve AVSR results but still underperforms our proposed approach by a large margin.

**Performance on Single-Modality Inputs.** Furthermore, Table 5 presents the performance of our MIR-GAN on single-modality inputs. First, we observe that in all models using both modalities performs better than single modality, and the audio-only case achieves much better results than visual-only case, which shows the dominance of audio modality in AVSR task. Under two single-modality cases, our proposed MIR-GAN both achieves significant improvement over the base model, and the best performance outperforms or matches previous state-of-the-arts in both supervised and self-supervised settings (2.3% vs. 2.3%, 34.2% vs. 33.6%; 1.3% vs. 1.4%, 26.6% vs. 26.9%). Therefore, even with missing modality, our MIR-GAN can still refine effective modality-invariant representations to benefit the downstream speech recognition, which further verifies the generality of our approach.

**Conclusion**

In this paper, we propose MIR-GAN, an adversarial network to refine frame-level modality-invariant representations, and similar performance drops can be observed in absence of such strategy. Furthermore, we visualize the attention maps in Fig. 4 to show its effectiveness. The clear diagonals in (c) and (d) indicate the strong ability of MIM strategy to align modality-invariant and -specific representations, which enriches the contextual semantic information in the former.

**Table 4:** Comparison between MIR-GAN and utterance-level multimodal approaches on LRS3 benchmark.

<table>
<thead>
<tr>
<th>Method</th>
<th>WER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clean</td>
</tr>
<tr>
<td>Base Model</td>
<td>3.5</td>
</tr>
<tr>
<td>+ MCLNet (Hao et al., 2021)</td>
<td>3.4</td>
</tr>
<tr>
<td>+ VI-REID (Feng et al., 2019)</td>
<td>3.3</td>
</tr>
<tr>
<td>+ MISA (Hazarika et al., 2020)</td>
<td>3.3</td>
</tr>
<tr>
<td>MIR-GAN (ours)</td>
<td>2.8</td>
</tr>
</tbody>
</table>

Audio modality plays the dominant role in AVSR task.
<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Supervised</th>
<th>Self-Supervised</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AV</td>
<td>A</td>
</tr>
<tr>
<td>TM-seq2seq (2018a)</td>
<td>Transformer</td>
<td>7.2</td>
<td>8.3</td>
</tr>
<tr>
<td>EG-seq2seq (2020)</td>
<td>RNN</td>
<td>6.8</td>
<td>7.2</td>
</tr>
<tr>
<td>RNN-T (2019)</td>
<td>RNN</td>
<td>4.5</td>
<td>4.8</td>
</tr>
<tr>
<td>Hyb-Conformer (2021)</td>
<td>Conformer</td>
<td><strong>2.3</strong></td>
<td><strong>2.3</strong></td>
</tr>
<tr>
<td>Distill-Pretrain (2022)</td>
<td>Conformer</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>aV-HuBERT (2022b)</td>
<td>Transformer</td>
<td>1.4</td>
<td>1.5</td>
</tr>
<tr>
<td>u-HuBERT (2022)</td>
<td>Transformer</td>
<td>1.2</td>
<td>1.4</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>Base model</td>
<td>3.5</td>
<td>4.7</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>MIR-GAN</td>
<td>2.8</td>
<td>3.5</td>
</tr>
<tr>
<td><strong>Proposed (Supervised)</strong></td>
<td>Base model</td>
<td>2.5</td>
<td>3.0</td>
</tr>
<tr>
<td><strong>Proposed (Self-Supervised)</strong></td>
<td>MIR-GAN</td>
<td><strong>2.1</strong></td>
<td><strong>2.3</strong></td>
</tr>
</tbody>
</table>

Table 5: Performance on single-modality inputs with LRS3 benchmark. “AV”, “A” and “V” indicate the input modality during both finetuning and inference stages. The missing modality is replaced by zero embeddings.

representations for AVSR, which captures the commonality across modalities to ease the multimodal fusion process. MIR-GAN first learns modality-invariant representation with MIR generator, followed by a modality discriminator to strengthen its modality agnosticism via adversarial learning. Furthermore, we propose a mutual information maximization strategy to enrich its contextual semantic information. Finally, both modality-invariant and -specific representations are fused to provide a holistic view of multimodal data for downstream task. Experiments on public benchmarks show that our MIR-GAN achieves the state-of-the-art.

Limitations

The main novelty of our proposed MIR-GAN is refining frame-level modality-invariant representations via adversarial learning. It is promising to combine this approach with the popular self-supervised pre-training to learn unified multimodal representations. In this work, we only load pre-trained AV-HuBERT for the front-ends and speech recognition model, while the proposed modules (i.e., encoders, generator, discriminator) are still trained from scratch. In future, we may include the entire MIR-GAN into self-supervised learning scheme, together with the adversarial learning to refine better multimodal representations.

Ethics Statement

All the data used in this paper are publicly available and are used under the following five licenses: the Creative Commons BY-NC-ND 4.0 License and Creative Commons Attribution 4.0 International License, the TED Terms of Use, the YouTube’s Terms of Service, and the BBC’s Terms of Use. The data is collected from TED and BBC and contain thousands of speakers from a wide range of races. To protect the anonymity, only the mouth area of a speaker is visualized wherever used in the paper.

Acknowledgements

The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

References


Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022a. Learning audio-visual speech representation by masked multimodal cluster prediction. In International Conference on Learning Representations.


A.1 Datasets

**LRS3** (Afouras et al., 2018b) is currently the largest public sentence-level lip reading dataset, which contains over 400 hours of English video extracted from TED and TEDx talks on YouTube. The training data is divided into two parts: pretrain (403 hours) and trainval (30 hours), and both of them are transcribed at sentence level. The pretrain part differs from trainval in that the duration of its video clips are at a much wider range. Since there is no official development set provided, we randomly select 1,200 samples from trainval as validation set (~1 hour) for early stopping and hyper-parameter tuning. In addition, it provides a standard test set (0.9 hours) for evaluation.

**LRS2** (Chung et al., 2017) is a large-scale publicly available labeled audio-visual (A-V) datasets, which consists of 224 hours of video clips from BBC programs. The training data is divided into three parts: pretrain (195 hours), train (28 hours) and val (0.6 hours), which are all transcribed at sentence level. An official test set (0.5 hours) is provided for evaluation use.

A.2 Data Preprocessing

The data preprocessing for above two datasets follows the LRS3 preprocessing steps in prior work (Shi et al., 2022a). For the audio stream, we extract the 26-dimensional log filter-bank feature at a stride of 10 ms from input raw waveform.

---

6https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs3.html
7https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html
<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Criterion</th>
<th>Unlabeled data (hrs)</th>
<th>Labeled data (hrs)</th>
<th>DataAug</th>
<th>LM</th>
<th>WER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Supervised</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TM-seq2seq (2018a)</td>
<td>Transformer</td>
<td>S2S</td>
<td>-</td>
<td>1.519</td>
<td>✓</td>
<td>✓</td>
<td>8.5</td>
</tr>
<tr>
<td>TM-CTC (2018a)</td>
<td>Transformer</td>
<td>CTC</td>
<td>-</td>
<td>1.519</td>
<td>✓</td>
<td>✓</td>
<td>8.2</td>
</tr>
<tr>
<td>Hyb-RNN (2018)</td>
<td>RNN</td>
<td>S2S + CTC</td>
<td>-</td>
<td>397</td>
<td>✓</td>
<td>✓</td>
<td>7.0</td>
</tr>
<tr>
<td>LF-MMI TDNN (2020)</td>
<td>TDNN</td>
<td>LF-MMI</td>
<td>-</td>
<td>224</td>
<td>-</td>
<td>✓</td>
<td>5.9</td>
</tr>
<tr>
<td>Hyb-Conformer (2021)</td>
<td>Conformer</td>
<td>S2S + CTC</td>
<td>-</td>
<td>381</td>
<td>✓</td>
<td>✓</td>
<td>3.7</td>
</tr>
<tr>
<td><strong>Self-Supervised</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MoCo+wav2vec (2022)</td>
<td>Transformer</td>
<td>S2S + CTC</td>
<td>60,000</td>
<td>381</td>
<td>✓</td>
<td>-</td>
<td>2.6</td>
</tr>
<tr>
<td><strong>Proposed (Supervised)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>Base model</td>
<td>Transformer</td>
<td>-</td>
<td>224</td>
<td>✓</td>
<td>-</td>
<td>5.4</td>
</tr>
<tr>
<td>MIR-GAN</td>
<td>Base model</td>
<td>Transformer</td>
<td>S2S -</td>
<td>224</td>
<td>✓</td>
<td>-</td>
<td>3.9</td>
</tr>
<tr>
<td><strong>Proposed (Self-Supervised)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>Base model</td>
<td>Transformer</td>
<td>S2S -</td>
<td>1,759</td>
<td>✓</td>
<td>-</td>
<td>2.3</td>
</tr>
<tr>
<td>MIR-GAN</td>
<td>Base model</td>
<td>Transformer</td>
<td>S2S -</td>
<td>224</td>
<td>✓</td>
<td>-</td>
<td>2.2</td>
</tr>
</tbody>
</table>

Table 6: WER (%) of our MIR-GAN and prior works on LRS2 benchmark. “S2S” denotes sequence-to-sequence loss (Watanabe et al., 2017), “CTC” denotes CTC loss (Graves et al., 2006), “DataAug” denotes noise augmentation, “LM” denotes language model rescoring. The noisy test set is synthesized using MUSAN noise (Snyder et al., 2015).

For the video clips, we detect the 68 facial keypoints using dlib toolkit (King, 2009) and align the image frame to a reference face frame via affine transformation. Then, we convert the image frame to gray-scale and crop a 96×96 region-of-interest (ROI) centered on the detected mouth. During training, we randomly crop a 88×88 region from the whole ROI and flip it horizontally with a probability of 0.5. At inference time, the 88×88 ROI is centered cropped without horizontal flipping. To synchronize these two modalities, we stack each 4 neighboring acoustic frames to match the image frames that are sampled at 25Hz.

A.3 Model Settings

Front-ends. We introduce the modified ResNet-18 from prior work (Shi et al., 2022a) as visual front-end, where the first convolutional layer is replaced by a 3D convolutional layer with kernel size of 5×7×7. The visual feature is flattened into an 1D vector by spatial average pooling in the end. For audio front-end, we use one linear projection layer followed by layer normalization (Ba et al., 2016).

MIR-GAN. We build the MIR-GAN framework based on Transformer, where the embedding dimension/feed-forward dimension/attention heads in each Transformer layer are set to 1024/4096/16 respectively. In addition, we also employ Conformer as backbone, where the depth-wise convolution kernel size is set to 31. We use a dropout of $p = 0.1$ after the self-attention block within each Transformer layer, and each Transformer layer is dropped (Fan et al., 2019) at a rate of 0.1.

A.4 Data Augmentation

Following prior work (Shi et al., 2022b), we use many noise categories for data augmentation. We select the noise categories of “babble”, “music” and “natural” from MUSAN noise dataset (Snyder et al., 2015), and extract some “speech” noise samples from LRS3 dataset. All categories are divided into training, validation and test partitions. During training process, we randomly select one noise category and sample a noise clip from its training partition. Then, we randomly mix the sampled noise with input clean audio, at signal-to-noise ratio (SNR) of 0dB with a probability of 0.25. At inference time, we evaluate our model on clean and noisy test sets respectively. Specifically, the system performance on each noise type is evaluated separately, where the testing noise clips are added at five different SNR levels: $\{-10, -5, 0, 5, 10\}$ dB. At last, the testing results on different noise types and SNR levels will be averaged to obtain the final noisy WER result.

A.5 Training Details

Training. We follow the sequence-to-sequence finetuning configurations of AV-HuBERT (Shi et al., 2022b) to train our systems. We use Transformer decoder to decode the encoded features into unigram-based subword units (Kudo, 2018), where
the vocabulary size is set to 1000. The temperature \( \tau \) in Eq. 7 is set to 0.1, and the weighting parameters \( \lambda_{GAN} / \lambda_{MIM} \) in Eq. 8 are set to 0.01/0.005 respectively. The entire system is trained for 60K steps using Adam optimizer (Kingma and Ba, 2014), where the learning rate is warmed up to a peak of 0.001 for the first 20K updates and then linearly decayed. The finetuning process takes \(~1.4\) days on 4 NVIDIA-V100-32GB GPUs.

**Inference.** No language model is used during inference. We employ beam search for decoding, where the beam width and length penalty are set to 50 and 1 respectively. All the hyper-parameters in our systems are tuned on validation set. Since our experimental results are quite stable, a single run is performed for each reported result.

### A.6 Baselines

In this section, we describe the baselines for comparison.

- **TM-seq2seq** (Afouras et al., 2018a): TM-seq2seq proposes a Transformer-based AVSR system to model the A-V features separately and then attentively fuse them for decoding, and uses sequence-to-sequence loss (Watanabe et al., 2017) as training criterion.

- **TM-CTC** (Afouras et al., 2018a): TM-CTC shares the same architecture with TM-seq2seq, but uses CTC loss (Graves et al., 2006) as training criterion.

- **Hyb-RNN** (Petridis et al., 2018): Hyb-RNN proposes a RNN-based AVSR model with hybrid seq2seq/CTC loss (Watanabe et al., 2017), where the A-V features are encoded separately and then concatenated for decoding.

- **RNN-T** (Makino et al., 2019): RNN-T adopts the popular recurrent neural network transducer (Graves, 2012) for AVSR task, where the audio and visual features are concatenated before fed into the encoder.

- **EG-seq2seq** (Xu et al., 2020): EG-seq2seq builds a joint audio enhancement and multimodal speech recognition system based on the element-wise attention gated recurrent unit (Zhang et al., 2019), where the A-V features are concatenated before decoding.

- **LF-MMI TDNN** (Yu et al., 2020): LF-MMI TDNN proposes a joint audio-visual speech separation and recognition system based on time-delay neural network (TDNN), where the A-V features are concatenated before fed into the recognition network.

- **Hyb-Conformer** (Ma et al., 2021): Hyb-Conformer proposes a Conformer-based (Gulati et al., 2020) AVSR system with hybrid seq2seq/CTC loss, where the A-V input streams are first encoded separately and then concatenated for decoding.

- **MoCo+wav2vec** (Pan et al., 2022): MoCo+wav2vec employs self-supervised pre-trained audio and visual front-ends, *i.e.*, wav2vec 2.0 (Baevski et al., 2020) and MoCo v2 (Chen et al., 2020), to generate better audio-visual features for fusion and decoding.

- **AV-HuBERT** (Shi et al., 2022a,b): AV-HuBERT employs self-supervised learning to capture deep A-V contextual information, where the A-V features are masked and concatenated before fed into Transformer encoder to calculate masked-prediction loss for pre-training, and cross-entropy based sequence-to-sequence loss is used for finetuning.

- **u-HuBERT** (Hsu and Shi, 2022): u-HuBERT extends AV-HuBERT to a unified framework of audio-visual and audio-only pre-training.

- **Distill-Pretrain** (Ma et al., 2022): Distill-Pretrain proposes a Conformer-based VSR framework with additional distillation from pre-trained ASR and VSR models.