

Multi-modality Associative Bridging through Memory: Speech Sound Recollected from Face Video

Minsu Kim* Joanna Hong* Se Jin Park Yong Man Ro†
Image and Video Systems Lab, KAIST, South Korea
{ms.k, joanna2587, jinny960812, ymro}@kaist.ac.kr

Abstract

In this paper, we introduce a novel audio-visual multi-modal bridging framework that can utilize both audio and visual information, even with uni-modal inputs. We exploit a memory network that stores source (i.e., visual) and target (i.e., audio) modal representations, where source modal representation is what we are given, and target modal representations are what we want to obtain from the memory network. We then construct an associative bridge between source and target memories that considers the inter-relationship between the two memories. By learning the interrelationship through the associative bridge, the proposed bridging framework is able to obtain the target modal representations inside the memory network, even with the source modal input only, and it provides rich information for its downstream tasks. We apply the proposed framework to two tasks: lip reading and speech reconstruction from silent video. Through the proposed associative bridge and modality-specific memories, each task knowledge is enriched with the recalled audio context, achieving state-of-the-art performance. We also verify that the associative bridge properly relates the source and target memories.

1. Introduction

Recently, many studies are dealing with diverse information from multiple sources finding relationships among them [40]. Especially, deep learning based multi-modal learning has drawn big attention with its powerful performance. While the classic approaches [20, 5, 52, 53] need to design each modal feature manually, using Deep Neural Networks (DNNs) has the advantage of automatically learning meaningful representation from each modality. Many applications including action recognition [15, 25], object detection [11], and image/text retrieval [64] have shown the effectiveness of multi-modal learning through DNNs by analyzing a phenomenon in multi-view.

*Both authors have contributed equally to this work.

†Corresponding author

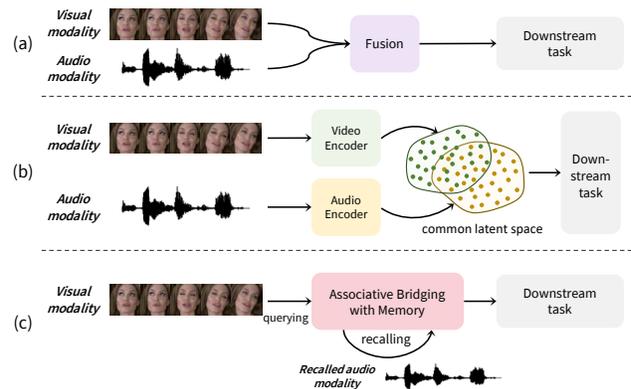


Figure 1. Illustration of audio-visual multi-modal learning. (a) Fusion of two modalities. (b) Learning from a common latent space of two modalities. (c) The proposed framework provides an associative bridge between two modalities through memory. The audio (i.e., target) modality is recalled from memory by querying the visual (i.e., source) modality. Then, both the visual and the recalled audio modalities are utilized for a downstream task.

Audio-visual data is one of the main ingredients for multi-modal applications such as synchronization [7, 8], speech recognition [1, 38], and speech reconstruction from silent video [39, 3]. Along with the rapid increase of the demands for audio-visual applications, research efforts on how to efficiently handle audio-visual data have been made. There are two main streams on handling audio-visual data. First is to extract features from the two modalities and fuse them to achieve complementary effect, as shown in Fig. 1 (a). Such researches [38, 1, 36] try to find the most suitable architecture of DNNs to fuse the modalities. Commonly used methods are early fusion, late fusion, and intermediate fusion. These fusion methods are known to be simple, yet effectively improve the performance of a given task. However, since both modalities are necessary for the fusion, these methods cannot work when one of the modalities is missing. Second is finding a common hidden representation of two modalities by training DNNs (Fig. 1 (b)). Different from the first method, it can utilize the shared information of both modalities from the learned cross-modal

representation with uni-modal inputs. This can be achieved by finding the common latent space of different modalities using metric learning [7, 8] or resembling the other modality which contains rich information for a given task using knowledge distillation [63]. However, reducing the heterogeneity gap [21], induced by inconsistent distribution of different modalities, is still considered as a challenging problem [19, 37].

In this paper, we propose a novel multi-modal bridging framework, especially in audio speech modality and visual face modality. The proposed framework brings the advantages of the two aforementioned audio-visual multi-modal learning methods, while alleviating the problems that each method contains. That is, it can obtain both audio and visual contexts during inference even when the uni-modal input is provided only. This gives explicit complementary knowledge with the multi-modal information to uni-modal tasks which could suffer from information insufficiency. Furthermore, our work can be free from finding a common representation of different modalities, as shown in Fig.1(c).

To this end, we propose to handle the audio-visual data through memory network [55, 32] which contains two modality-specific memories: source-key memory and target-value memory. Each memory stores visual and audio features arranged in pairs, respectively. Then, an associative bridge is constructed between the two modality-specific memories, to access the target-value memory by querying the source-key memory with source modal representation. Thus, when one modality (*i.e.*, source) is given, the proposed framework can recall the other saved modality (*i.e.*, target) from target-value memory through the associative bridge. This enables it to complement the information of uni-modal input with the recalled target modal information. Therefore, we can enrich the task-solving ability of a downstream task. The proposed framework is verified on two applications using audio-visual data: lip reading, and speech reconstruction from silent video by using visual modality as source modality and audio modality as target modality.

In summary, the major contributions of this paper are as follows:

- We propose a novel audio-visual multi-modal bridging framework that enables it to utilize the information of multi-modality (*i.e.*, audio and visual modalities) with uni-modal (*i.e.*, visual) input during inference.
- We verify the effectiveness of the proposed framework on two applications: lip reading and speech reconstruction from silent video and achieve state-of-the-art performances. Moreover, we visualize that the associative bridge adequately relates the source and target memories.
- Through the proposed modality-specific memory operation (*i.e.*, querying by source modality and recalling

target modality), it does not need to find a common latent space of different modalities. We analyze it by comparing the proposed framework with the methods finding a common latent space of multi-modal data.

2. Related Work

2.1. Multi-modal learning with audio-visual data

Audio-visual multi-modal learning is one of the active research areas. There are two categories of audio-visual multi-modal learning using DNNs: fusion and finding a common latent space of cross-modal representation. The fusion methods [48, 33, 10] aim to exploit the complementary information of different modalities and achieve high performance compared to the uni-modal methods. They try to find the best fusion architecture of a given task [1, 34, 36, 28]. However, as the fusion methods receive all modalities as inputs, they could not properly work if one of them is not available. The learning methods finding a common latent space from multi-modal data [35, 19, 4, 24] aim to reduce the heterogeneity gap between the two modalities. Several works [7, 8, 14] have proposed metric learning methods and adversarial learning methods to find the common representation. Other works [63, 2] have proposed to learn from superior modality for a given task using knowledge distillation [18] which guides the learned feature to resemble the superior modal feature. Although finding a shared latent space or guiding one modal representation to resemble the other has the advantage of using the common information between the two modalities with uni-modal inputs, reducing the heterogeneity gap between the multi-modal data is considered as a challenging problem [64, 37].

In this paper, we try to not only take the advantages of both methods, but also alleviate the problems of each method. We propose to handle the audio-visual data using two modality-specific memory networks connected with an associative bridge. During inference, the proposed framework can exploit both source and the recalled target modal contexts even when the input is uni-modal. Moreover, since each modality works on its corresponding modality-specific module, we can bypass the difficulty of finding a shared latent space.

2.2. Memory network

Memory network is a scheme to augment neural networks using external memory [55, 46]. They have shown the effectiveness of memory network on modeling long-term dependencies in sequential data [29]. Miller *et al.* [32] introduce key-value paired memory structure where key memory is firstly used to address relevant memories with respect to a query, extracting addressed values from the value memory. We utilize the key-value memory network [32], where the key memory is for saving the source

modal features, and the value memory is for saving the target modal features. Thus, we can access both source and target modal contexts by recalling the saved target modal feature from the value memory when only source modality is available.

The memory network is also used in multi-modal modeling. Song *et al.* [44] introduce a cross-modal memory network for cross-modal retrieval. Huang *et al.* [22] propose an aligned cross-modal memory network for few-shot image and sentence matching. Using a shared memory, they encode memory-enhanced features, which will be used for image/text matching. Distinct from the previous methods, our proposed framework uses modality-specific memory network where source-key memory saves the source modality and target-value memory saves the target modality.

2.3. Lip reading

Lip reading is a task that recognizes speech as text from lip movements. Chung *et al.* [6] propose word-level audio-visual corpus data and a baseline architecture. The performance of word-level lip reading is significantly improved with the architecture [45, 38] of a 3D convolution layer, a ResNet-34, and Bi-RNNs. Some works [54, 56] use both optical flow and video frames to capture fine-grained motion. Xu *et al.* [57] suggest a pseudo-3D CNN for the frontend which is more efficient compared to vanilla 3D CNN. Zhang *et al.* [61] show that the lip reading can be made beyond the lips by utilizing entire face as inputs. Martinez *et al.* [31] improve the backend by changing the Bi-RNN into multi-scale temporal CNN.

It is widely known that the audio modality has superior knowledge for speech recognition than the visual modality by showing outstanding performance. In this paper, we try to complement the lip visual information by recalling the speech audio information from the proposed multi-modal bridging framework.

2.4. Speech reconstruction from silent video

Speech reconstruction from silent video aims to generate acoustic speech signal from silent talking face video. Ephrat *et al.* [13, 12] firstly generate speech using CNN and they improve it with a two tower CNN-based encoder-decoder architecture whose inputs are both optical flow and video frames. Akbari *et al.* [3] propose to pretrain an auto-encoder to reconstruct the speech, whose decoder part is used to generate the speech from a face video. Vougioukas *et al.* [50] propose GAN based approach which maps video directly to audio waveform. Prajwal *et al.* [39] attempt to learn on unconstrained single-speaker dataset. They present a model that consists of stacked 3D convolutions and an attention-based speech decoder, formulating the task as a sequence-to-sequence problem.

The speech reconstruction from silent video is consid-

ered as a challenging problem, due to the information insufficiency of face visual movement to fully represent the speech audio. We try to provide complementary information with the recalled audio representation through the proposed associative bridge with memory, and enhance its performance. With both the visual and the recalled audio contexts, we can generate high-quality speech in both speaker-dependent and speaker-independent settings.

3. Multi-modality Associative Bridging

The main objective of the proposed framework is to recall the target modal representation with only source modal inputs. To this end, (1) each modality-specific memory is guided to save the representative features of each modality, and (2) an associative bridge is constructed which enables it to recall the target modal representation by querying the source-key memory with source modal feature. As shown in Fig.2, the proposed multi-modality associative bridging framework is composed of two modality-specific memory networks: source-key memory $\mathbf{M}_{src} \in \mathbb{R}^{N \times C}$ and target-value memory $\mathbf{M}_{tgt} \in \mathbb{R}^{N \times D}$, where N represents the number of memory slots, and C and D are the dimension of each modal feature, respectively. From the following subsections, we will describe the details of the proposed framework with examples of visual modality as source modality and the audio modality as target modality.

3.1. Embedding modality-specific representations

Each memory network inside the proposed framework saves generic representations of each modality. The generic visual and audio representations are produced from the respective modality-specific deep embedding modules. The visual (*i.e.*, source modal) representation $f_{src} \in \mathbb{R}^{T \times C}$ is extracted by using spatio-temporal CNN that captures both spatial and temporal information, and the audio (*i.e.*, target modal) representation $f_{tgt} \in \mathbb{R}^{T \times D}$ is embedded from 2D CNN whose input is preprocessed mel-spectrogram from raw audio signal, where T represents the temporal length of each representation. Since the paired audio-video inputs are synchronized in time, the two embedding modules can be designed to output the same temporal length.

3.2. Addressing modality-specific memory

Based on the modality-specific representations, we firstly introduce how the source and target addressing vectors are formulated. The addressing vector refers to the guidance that determines where to assign weights on memory slots for a given query. Suppose that the source modal representation f_{src} is given as a query, then the cosine similarity with source-key memory \mathbf{M}_{src} is obtained,

$$s_{src}^{i,j} = \frac{\mathbf{M}_{src}^i \cdot f_{src}^j}{\|\mathbf{M}_{src}^i\|_2 \cdot \|f_{src}^j\|_2}, \quad (1)$$

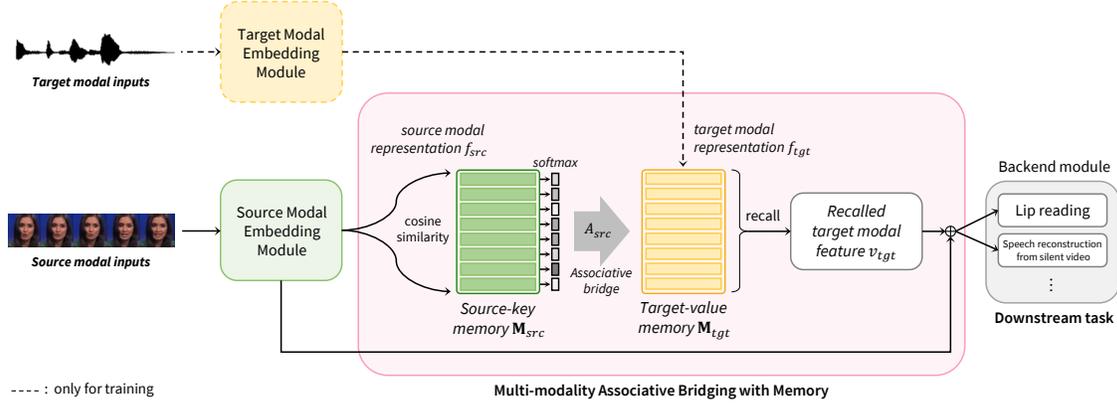


Figure 2. Overview of the proposed multi-modal bridging framework with an example of visual modality as a source and the audio modality as a target. The source-key memory is for saving source modal feature, and the target-value memory is for memorizing the target modal representations.

where $s_{src}^{i,j}$ represents the cosine similarity between i -th memory slot of source-key memory and source modal feature in j -th temporal step. Next, the relevance probability is obtained using Softmax function as follows,

$$\alpha_{src}^{i,j} = \frac{\exp(r \cdot s_{src}^{i,j})}{\sum_{k=1}^N \exp(r \cdot s_{src}^{k,j})}, \quad (2)$$

where r is a scaling factor for similarity. By calculating the probability over the entire memory slot, the source addressing vector for the j -th temporal step $A_{src}^j = \{\alpha_{src}^{1,j}, \alpha_{src}^{2,j}, \dots, \alpha_{src}^{N,j}\}$ can be obtained.

The same procedure is applied for target modal representation f_{tgt} and target-value memory M_{tgt} to produce the target addressing vector, $A_{tgt}^j = \{\alpha_{tgt}^{1,j}, \alpha_{tgt}^{2,j}, \dots, \alpha_{tgt}^{N,j}\}$ of j -th temporal step. The addressing vectors will be utilized in recalling the saved representations inside memory and connecting the two modality-specific memories, in the following subsections.

3.3. Memorizing the target modal representations

The obtained target addressing vector A_{tgt}^j is to correctly match the target-value memory M_{tgt} for reconstructing target representation \hat{f}_{tgt}^j . To do so, the target-value memory M_{tgt} is trained to memorize the proper target modal representation f_{tgt}^j . We firstly obtain the reconstructed target representation \hat{f}_{tgt}^j as follows,

$$\hat{f}_{tgt}^j = A_{tgt}^j \cdot M_{tgt}. \quad (3)$$

Then, we design the reconstruction loss function to guide the target-value memory M_{tgt} to save the proper representation. We minimize the Euclidean distance between the target representation and the reconstructed representation,

$$\mathcal{L}_{save} = \mathbb{E}_j[\|f_{tgt}^j - \hat{f}_{tgt}^j\|_2^2]. \quad (4)$$

With the saving loss, the target-value memory M_{tgt} saves the representative features of the target modality. Therefore, the recalled target modal representation \hat{f}_{tgt}^j from target-value memory M_{tgt} is able to represent the original target modal representation f_{tgt} .

3.4. Bridging source and target memories

To recall the target modal representation from the target-value memory by using the source-key memory and source modal inputs, we construct an associative bridge between the two modality-specific memories. Specifically, the source-key memory is utilized to provide the bridge between source and target modalities in the form of the source addressing vector. That is, through the source addressing vector A_{src}^j , the corresponding saved target representation is recalled. To achieve this, the source addressing vector A_{src}^j is guided to match to the target addressing vector A_{tgt}^j with the following bridging loss,

$$\mathcal{L}_{bridge} = \mathbb{E}_j[D_{KL}(A_{tgt}^j || A_{src}^j)], \quad (5)$$

where $D_{KL}(\cdot)$ represents Kullback–Leibler divergence [27]. With the bridging loss, the source-key memory saves the source modal representations in the same location, where the target-value memory saves the corresponding target modal features. Therefore, when a source modal representation is given, the source-key memory provides the location information of the corresponding saved target modal representation in the target-value memory, using the source addressing vector.

3.5. Applying for downstream tasks

Through the associative bridge and the modality-specific memories, we can obtain the recalled target modal feature v_{tgt} by using source addressing vector A_{src} as follows,

$$v_{tgt}^j = A_{src}^j \cdot M_{tgt}. \quad (6)$$

Here, the target modal feature v_{tgt} is recalled by querying the source-key memory \mathbf{M}_{src} with the source modal representation f_{src} . Thus, we do not need the target modal inputs for recalling the target modal feature. Then, we can apply the recalled target modal feature for a downstream task in addition to the source modality, improving task performance by exploiting the complementary information.

3.6. End-to-End training

The proposed framework is trainable in an end-to-end manner, including the modality-specific embedding modules, memory networks, and the downstream sub-networks. To this end, the following task loss is applied,

$$\mathcal{L}_{task} = g(h(f_{src} \oplus v_{tgt}); y) + g(h(f_{src} \oplus f_{tgt}); y), \quad (7)$$

where $g(\cdot)$ is a loss function corresponding to the downstream task, $h(\cdot)$ is a fusion layer such as a linear layer, y represents label, and \oplus represents concatenation. The first term of the loss function is related to the performance on a given task that utilizes both the source and the recalled target modalities. The second term guarantees that the target modal embedding module learns the meaningful representations that will be saved into target-value memory in an end-to-end manner.

Finally, the total loss function is defined as a sum of the all loss functions,

$$\mathcal{L}_{total} = \mathcal{L}_{save} + \mathcal{L}_{bridge} + \mathcal{L}_{task}. \quad (8)$$

The pseudo code for training the proposed framework is shown at Algorithm 1.

4. Experiments

The main strength of the proposed audio-visual bridging framework is that it is possible to use multi-modal representation even if only one modal input is available. Therefore, we can enhance the uni-modal downstream tasks by exploiting complementary information from the recalled modal features. We show the effectiveness of the proposed framework on two applications, lip reading and speech reconstruction from silent video, each of which takes visual modality as an input. Therefore, visual modality is utilized as a source modality and audio modality is used as a target modality.

4.1. Application 1: Lip reading

Lip reading is a task that recognizes speech by solely depending on lip movements. We apply the proposed multi-modal bridging framework to the lip reading to complement the visual context by bringing superior knowledge of the audio through the associative bridge and to enhance the performance.

Algorithm 1 Training algorithm of the proposed framework

- 1: **Inputs:** The training pairs of source and target modal inputs (X_{src}, X_{tgt}) and label y , where $X_{src} = \{x_{src}^l\}_{l=1}^L$, $X_{tgt} = \{x_{tgt}^s\}_{s=1}^S$. The learning rate η .
- 2: **Output:** The optimized parameters of the network Φ

- 3: Randomly initialize parameters of the network Φ
- 4: **for** each iteration **do**
- 5: $f_{src} = \{f_{src}^j\}_{j=1}^T = \text{Source_embed}(X_{src})$
- 6: $f_{tgt} = \{f_{tgt}^j\}_{j=1}^T = \text{Target_embed}(X_{tgt})$
- 7: **for** $j = 1, 2, \dots, T$ **do**
- 8: $A_{src}^j = \text{Softmax}(r \cdot \text{CosineSim}(\mathbf{M}_{src}, f_{src}^j))$
- 9: $A_{tgt}^j = \text{Softmax}(r \cdot \text{CosineSim}(\mathbf{M}_{tgt}, f_{tgt}^j))$
- 10: $\hat{f}_{tgt}^j = A_{tgt}^j \cdot \mathbf{M}_{tgt}$
- 11: $v_{tgt}^j = A_{src}^j \cdot \mathbf{M}_{tgt}$
- 12: **end for**
- 13: $\mathcal{L}_{save} = \sum_{j=1}^T \|f_{tgt}^j - \hat{f}_{tgt}^j\|_2^2$
- 14: $\mathcal{L}_{bridge} = \sum_{j=1}^T D_{KL}(A_{tgt}^j \| A_{src}^j)$
- 15: $\mathcal{L}_{task} = g(h(f_{src} \oplus v_{tgt}); y) + g(h(f_{src} \oplus f_{tgt}); y)$
- 16: $\mathcal{L}_{tot} = \mathcal{L}_{save}/T + \mathcal{L}_{bridge}/T + \mathcal{L}_{task}$
- 17: Update $\Phi \leftarrow \Phi - \eta \nabla_{\Phi} \mathcal{L}_{tot}$
- 18: **end for**

4.1.1 Dataset

We utilize two public benchmark databases for word-level lip reading, LRW [6] and LRW-1000 [60]. Both datasets are composed of 25 fps video and 16kHz audio.

LRW [6] is a large-scale word-level English audio-visual dataset. It includes 500 words with a maximum of 1,000 training videos each. For the preprocessing, the video is cropped into 136×136 size centered at the lip, resized into 112×112 , and converted into grayscale. For the data augmentation, we use random horizontal flipping and random erasing for all frames in a video consistently. The audio is preprocessed using window size of 400, hop size of 160, and 80 mel-filter banks. Thus, the preprocessed mel-spectrogram has 100 fps with 80 spectral dimensional features. We use SGD optimizer, batch size of 320, and initial learning rate of 0.03.

LRW-1000 [60] is Mandarin words audio-visual dataset. It consists of 718,018 video samples with 1,000 word classes. The same preprocessing and data augmentation are applied as in LRW preprocessing except for cropping because the dataset is already cropped. Moreover, since the audio provided from the dataset is longer than the word boundary by 0.4-sec, we use the video as the same length as the audio. We use Adam [26] optimizer, batch size of 60, and initial learning rate of 0.0001.

4.1.2 Architecture

For the baseline architecture, we follow the typical architecture [38, 45] whose visual embedding module consists of one 3D convolution layer and ResNet-18 [17], and back-end module is composed of 2 layered Bi-GRU [42]. We

design the audio embedding module to output the same sequence length as that of the visual embedding module. For the task loss $g(\cdot)$, cross entropy loss is applied. The details of the network architecture can be found in supplementary.

4.1.3 Results

In order to verify the effectiveness of the proposed multi-modal bridging framework on complementing the visual modality with recalled audio modality, we compare the word-level lip reading using only visual modal inputs on benchmark datasets with the state-of-the-art methods. Table 1 shows the overall lip reading performances on LRW and LRW-1000 datasets. Our proposed framework achieves the highest accuracies among the previous approaches on both datasets. Especially for LRW-1000, which is known to be a difficult dataset due to unbalanced training samples, the proposed method attains a large improvement of 5.58% from the previous state-of-the-art method [61]. From this result, we can confirm that the proposed framework is even more effective for the difficult task with the ability of complementing the insufficient visual information with the recalled audio. Moreover, since our multi-modal associative bridging framework is not dependent on the downstream architecture, deep architecture such as temporal CNN can be adopted to the proposed method to improve word prediction performance.

We also conduct an ablation study with four different models for each language (*i.e.*, $N=0, 44, 88, 132$ for English and $N=0, 56, 112, 168$ for Mandarin) to examine the effect of the number of memory slots. The ablation results on memory slot size are reported in supplementary material. For LRW, the best word accuracy of 85.41% is achieved when $N=88$. The proposed framework improves the baseline with a margin of 1.27%. For LRW-1000, the best word accuracy is 50.82% when $N=112$ by improving the baseline performance with 5.89%. The proposed framework improves the performance regardless of the number of memory slots from the baseline in both languages.

By employing the recalled audio feature as complementary information of the visual context, the proposed framework successfully refines the word prediction achieving state-of-the-art performance.

4.2. Application 2: Speech reconstruction from silent video

Speech reconstruction from silent video is a task of inferring the speech audio signal by watching the facial video. To demonstrate the effectiveness of the proposed multi-modal bridging framework, we apply the proposed framework to the speech reconstruction from silent video task to provide the recalled audio context in an early stage of decoding for generating a high quality speech.

Method	LRW	LRW-1000
Yang <i>et al.</i> [60]	83.0	38.19
Multi-Grained [51]	83.3	36.91
PCPG [30]	83.5	38.70
Deformation Flow [56]	84.1	41.93
MI Maximization [62]	84.4	38.79
Face Cutout [61]	85.0	45.24
MS-TCN [31]	85.3	41.40
Proposed Method	85.4	50.82

Table 1. Lip reading word accuracy comparison with visual modal inputs on LRW and LRW-1000 dataset.

4.2.1 Dataset

GRID dataset [9] contains short English phrases with 6 words from predefined dictionary. The video and audio are sampled with rate of 25fps and 16kHz, respectively. Following [50, 39], subjects 1, 2, 4, and 29 are taken for speaker-dependent task. For the speaker-independent setting, we follow the same split as [50] which uses 15 subjects for training, 5 for validation, and 5 for testing. For the preprocessing, the face is detected, cropped and resized into 96×96 size. The audio is preprocessed with window size of 800, hop size of 160, and 80 mel-filter banks, becoming 80-dimensional mel-spectrogram in 100 fps. We use Adam optimizer, batch size of 64, and initial learning rate of 0.001.

4.2.2 Architecture

For the baseline architecture, we follow the state-of-the-art method [39] whose visual embedding module is composed of 3D CNN and Bi-LSTM. We adopt the backend module as the decoder part of Tacotron2 [43]. We utilize the same architecture of audio embedding module as lip reading experiment except for additional one convolution layer with kernel size of 5 before the Residual block. We adopt Griffin-Lim [16] algorithm for audio waveform conversion. For the task loss $g(\cdot)$, L1 distance loss is applied. More details of the network architecture can be found in supplementary material.

4.2.3 Results

We use three standard speech quality metrics for quantitative evaluation: STOI [47], ESTOI [23], and PESQ [41]. Table 2 shows the performance comparison on GRID dataset in speaker-dependent setting. We report the average test scores for 4 speakers with the same setting of the previous works [3, 50, 12, 39, 58]. The table clearly indicates that our model outperforms previous methods, including state-of-the-art performance. These improvements are from recalling the audio representations in the early stage of the backend which enables it to refine the generated mel-spectrogram.

Moreover, we ask 25 human participants to rate the Naturalness and Intelligibility. Naturalness is evaluating how

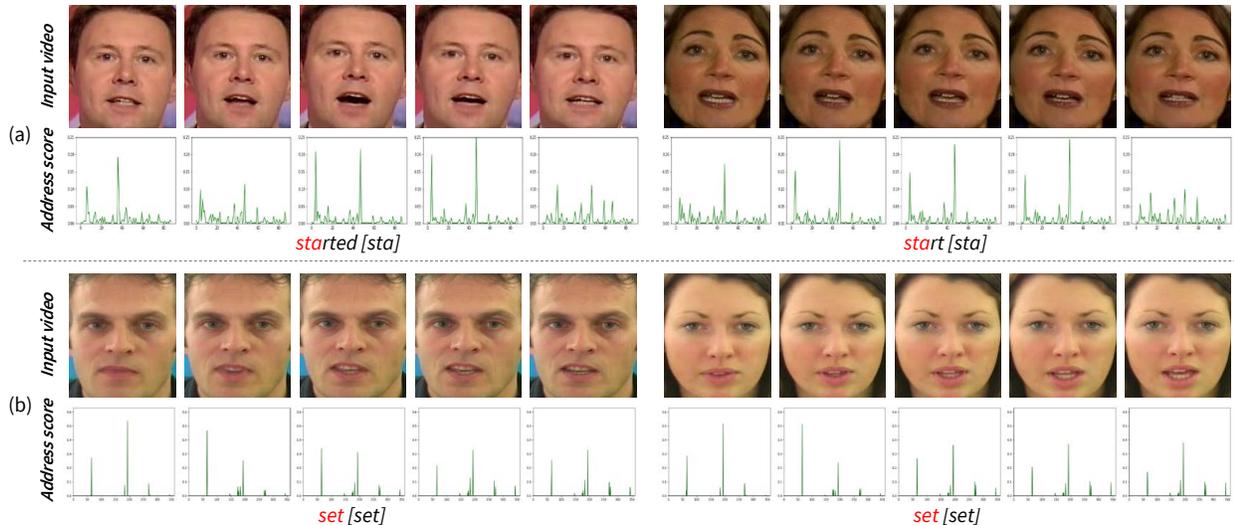


Figure 3. Face video clips (source modality) and corresponding addressing vectors for recalling audio modality (target modality) from learned representations inside memory: (a) results from lip reading and (b) results from speech reconstruction from silent video.

Method	STOI	ESTOI	PESQ
Vid2Speech [13]	0.491	0.335	1.734
Lip2AudSpec [3]	0.513	0.352	1.673
Vougioukas <i>et al.</i> [50]	0.564	0.361	1.684
Ephrat <i>et al.</i> [12]	0.659	0.376	1.825
Lip2Wav [39]	0.731	0.535	1.772
Yadav <i>et al.</i> [58]	0.724	0.540	1.932
Proposed Method	0.738	0.579	1.984

Table 2. Performance of speech reconstruction comparison with visual modal inputs in a speaker-dependent setting on GRID.

natural the synthetic speech is compared to the actual human voice, and intelligibility is how clearly the words sound in the synthetic speech compared to the actual transcription. 6 samples of generated speech for each of 4 speakers of GRID are used. The human subjective evaluation results are reported at Table 3. Compared to the previous works [13, 39], the proposed method achieves better scores on both Naturalness and Intelligibility. Moreover, with WaveNet [59] vocoder instead of Griffin-Lim, we can improve the scores as close to that of the ground truth. This indicates the reconstructed mel-spectrogram is of high-quality so that we can further improve the audio quality by using the state-of-the-art vocoder.

We also conduct an experiment on the speaker-independent setting, which is known to be a complex setting, of the GRID dataset to verify the effectiveness of the proposed method. As shown in Table 4, compared to [50, 39], the proposed framework achieves the highest performance. It can be inferred that even in a complex setting, the proposed framework can achieve meaningful outcomes by bringing the additional information through the associative bridge and memory. We visualize the examples of the generated mel-spectrogram in supplementary material.

Method	Naturalness	Intelligibility
Vid2Speech [13]	1.31 \pm 0.24	1.42 \pm 0.23
Lip2Wav [39]	2.83 \pm 0.21	2.94 \pm 0.19
Proposed Method	2.93 \pm 0.21	3.56 \pm 0.19
Proposed Method (+WaveNet vocoder [59])	4.37 \pm 0.16	4.27 \pm 0.14
Ground Truth	4.62 \pm 0.13	4.57 \pm 0.14

Table 3. Mean opinion scores for human evaluation on GRID.

Method	STOI	ESTOI	PESQ
Vougioukas <i>et al.</i> [50]	0.445	-	1.240
Lip2Wav [39]	0.565	0.279	1.279
Proposed Method	0.600	0.315	1.332

Table 4. Performance of speech reconstruction comparison with visual modal inputs on the speaker-independent setting on GRID.

We conduct an ablation study on different memory slot size, which is shown in supplementary material. It shows the best scores of 0.738 STOI, 0.579 ESTOI, and 1.984 PESQ when $N=150$. Moreover, the performance of the proposed framework improves regardless of the number of memory slots, which verifies its effectiveness.

4.3. Learned representation inside memory

In this section, we visualize the addressing vectors of both lip reading and speech reconstruction model in speaker-independent setting. Fig.3 (a) shows the video clips of LRW dataset with consecutive 5 frames and the corresponding addressing vectors of lip reading model. From the addressing vectors of different speakers speaking the same pronunciation, we observe the similar tendency of the addressing vectors. For example, when the face video is saying “sta” in words *started* and *start*, similar memory slots

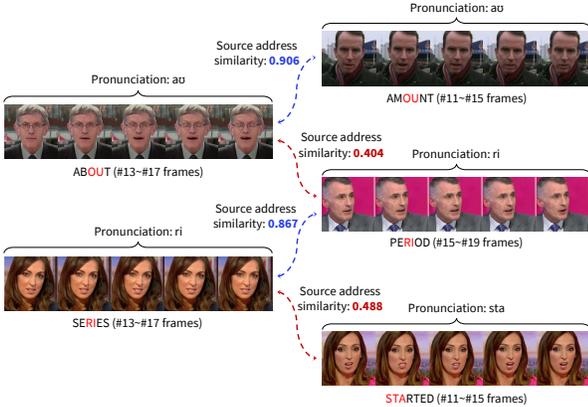


Figure 4. Examples of similarity between memory addressing vectors of different video clips in LRW. Note source addressing vector is for bridging video and audio modal features in memory.

are highly addressed. The same tendency can be observed in the speech reconstruction model shown in Fig.3 (b). This shows that source-key memory consistently finds the corresponding saved audio location in the target-value memory by using the talking face video clips as a query, which means the associative bridge is meaningfully constructed.

In addition, we compare addressing vectors of facial video clips with different pronunciations. Fig.4 shows the consecutive video frames with its corresponding pronunciation, and the comparison results. We can observe that the source addressing vectors of saying similar pronunciation have high similarity, while differently pronouncing videos have low similarity. For example, video clips of pronunciation “au” of word *about* and *amount* have 0.906 cosine similarity. In contrast, the similarity between “ri” of word *period* and “au” of word *about* is low with 0.404.

4.4. Comparison with methods finding a common latent space of multi-modality

We examine that the proposed framework can bypass the difficulty of finding a common representation of different modalities while bridging them. We compare the performance of word-level lip reading with the previous multi-modal learning methods that can exploit shared information of audio-visual modalities with uni-modal inference input by finding a common latent space. We build two multi-modal adaptation methods: cross-modal adaptation method [7] and knowledge distillation method [18]. The first is pre-trained to synchronize the audio-visual modalities, and then trained for lip reading. The second method is additionally trained so that the features from lip reading model resemble the features from the automatic speech recognition model.

We show the word-level lip reading accuracies on LRW dataset in Table 5. By utilizing multi-modality with visual modal inputs only, all of the methods show the performance improvements from the baseline, and the proposed framework achieves the best performance. The comparison shows

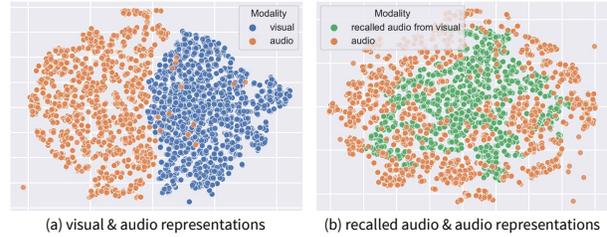


Figure 5. t-SNE [49] visualization of learned representation of (a) visual and audio modality, and (b) the recalled audio from visual modality and the actual audio modality.

Method	Baseline	Cross-modal Adaptation [7]	Knowledge Distillation [18]	Proposed Method
ACC(%)	84.14	84.20	84.50	85.41

Table 5. Lip reading word accuracy comparison with learning methods of finding a common representation of multi-modality.

the efficiency of the proposed framework, where it does not need to find a common latent space of two modalities by dealing with each modality in a modality-specific memory.

Lastly, we visualize the representations of visual modality, audio modality, and recalled audio modality from visual modality, by mapping them into 2D space. Fig.5 shows t-SNE [49] visualization of learned representations of visual and audio modalities, and the recalled audio from visual modality and the actual audio modality. Since we handle each modality with modality-specific embedding module and memory, the two modalities have separate representations in the latent space (Fig.5 (a)). However, as Fig.5 (b) shows, the recalled audio from the visual modality through the associative bridge shares a representation similar to the audio modal representation. Thus, we can utilize both audio and visual contexts while maintaining their own modal representations. This visualization demonstrates that we can effectively bridging the multi-modal representations without suffering from the cross-modal adaptation by dealing with each modality in modality-specific modules.

5. Conclusion

In this paper, we have introduced the multi-modality associative bridging framework that connects both audio and visual context through source-key memory and target-value memory. Thus, it can utilize both audio and visual information even if only one modality is available. We have verified the effectiveness of the proposed framework on two applications: lip reading and speech reconstruction from silent video, and achieved state-of-the-art performances. Furthermore, we have shown that the proposed framework can bridge the two modalities while maintaining separate latent space for each.

References

- [1] 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. 1, 2
- [2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Asr is all you need: Cross-modal distillation for lip reading. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2143–2147. IEEE, 2020. 2
- [3] Hassan Akbari, Himani Arora, Liangliang Cao, and Nima Mesgarani. Lip2audspec: Speech reconstruction from silent lip movements video. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2516–2520. IEEE, 2018. 1, 3, 6, 7
- [4] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013. 2
- [5] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, and Hua Zhang. Diversity-induced multi-view subspace clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–594, 2015. 1
- [6] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer, 2016. 3, 5
- [7] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016. 1, 2, 8
- [8] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3965–3969. IEEE, 2019. 1, 2
- [9] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006. 6
- [10] Stéphane Dupont and Juergen Luetttin. Audio-visual speech modeling for continuous speech recognition. *IEEE transactions on multimedia*, 2(3):141–151, 2000. 2
- [11] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust rgb-d object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687. IEEE, 2015. 1
- [12] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. Improved speech reconstruction from silent video. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 455–462, 2017. 3, 6, 7
- [13] Ariel Ephrat and Shmuel Peleg. Vid2speech: speech reconstruction from silent video. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5095–5099. IEEE, 2017. 3, 7
- [14] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16, 2014. 2
- [15] Zan Gao, Hai-Zhen Xuan, Hua Zhang, Shaohua Wan, and Kim-Kwang Raymond Choo. Adaptive fusion and category-level dictionary learning model for multiview human action recognition. *IEEE Internet of Things Journal*, 6(6):9280–9293, 2019. 1
- [16] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984. 6
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5
- [18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 8
- [19] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. Scalable deep multimodal learning for cross-modal retrieval. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 635–644, 2019. 2
- [20] Hsin-Chien Huang, Yung-Yu Chuang, and Chu-Song Chen. Affinity aggregation for spectral clustering. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 773–780. IEEE, 2012. 1
- [21] Xin Huang and Yuxin Peng. Deep cross-media knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8837–8846, 2018. 2
- [22] Yan Huang and Liang Wang. Acmm: Aligned cross-modal memory for few-shot image and sentence matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5774–5783, 2019. 3
- [23] Jesper Jensen and Cees H Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022, 2016. 6
- [24] Meina Kan, Shiguang Shan, and Xilin Chen. Multi-view deep network for cross-view classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4847–4855, 2016. 2
- [25] Jung Uk Kim, Sungjune Park, and Yong Man Ro. Uncertainty-guided cross-modal learning for robust multi-spectral pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 1
- [26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [27] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 4
- [28] Jong-Seok Lee and Cheol Hoon Park. Robust audio-visual speech recognition based on late integration. *IEEE Transactions on Multimedia*, 10(5):767–779, 2008. 2
- [29] Sangmin Lee, Hak Gu Kim, Dae Hwi Choi, Hyung-Il Kim, and Yong Man Ro. Video prediction recalling long-term motion context via memory alignment learning. In *Proceedings*

- of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3054–3063, 2021. 2
- [30] Mingshuang Luo, Shuang Yang, Shiguang Shan, and Xilin Chen. Pseudo-convolutional policy gradient for sequence-to-sequence lip-reading. *arXiv preprint arXiv:2003.03983*, 2020. 6
- [31] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Lipreading using temporal convolutional networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6319–6323. IEEE, 2020. 3, 6
- [32] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. *arXiv preprint arXiv:1606.03126*, 2016. 2
- [33] Ara V Nefian, Luhong Liang, Xiaobo Pi, Liu Xiaoxiang, Crusoe Mao, and Kevin Murphy. A coupled hmm for audio-visual speech recognition. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages II–2013. IEEE, 2002. 2
- [34] Chalapathy Neti, Gerasimos Potamianos, Juergen Luettin, Iain Matthews, Herve Glotin, Dimitra Vergyri, June Sison, and Azad Mashari. Audio visual speech recognition. Technical report, IDIAP, 2000. 2
- [35] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011. 2
- [36] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G Okuno, and Tetsuya Ogata. Audio-visual speech recognition using deep learning. *Applied Intelligence*, 42(4):722–737, 2015. 1, 2
- [37] Yuxin Peng and Jinwei Qi. Cm-gans: Cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1):1–24, 2019. 2
- [38] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. End-to-end audiovisual speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6548–6552. IEEE, 2018. 1, 3, 5
- [39] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13805, 2020. 1, 3, 6, 7
- [40] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017. 1
- [41] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE, 2001. 6
- [42] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997. 5
- [43] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018. 6
- [44] Ge Song, Dong Wang, and Xiaoyang Tan. Deep memory network for cross-modal retrieval. *IEEE Transactions on Multimedia*, 21(5):1261–1275, 2018. 3
- [45] Themos Stafylakis and Georgios Tzimiropoulos. Combining residual networks with lstms for lipreading. *arXiv preprint arXiv:1703.04105*, 2017. 3, 5
- [46] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. *Advances in neural information processing systems*, 28:2440–2448, 2015. 2
- [47] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, pages 4214–4217. IEEE, 2010. 6
- [48] Fei Tao and Carlos Busso. End-to-end audiovisual speech recognition system with multitask learning. *IEEE Transactions on Multimedia*, 23:1–11, 2020. 2
- [49] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8
- [50] Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Video-driven speech reconstruction using generative adversarial networks. *arXiv preprint arXiv:1906.06301*, 2019. 3, 6, 7
- [51] Chenhao Wang. Multi-grained spatio-temporal modeling for lip-reading. *arXiv preprint arXiv:1908.11618*, 2019. 6
- [52] Meng Wang, Xian-Sheng Hua, Richang Hong, Jinhui Tang, Guo-Jun Qi, and Yan Song. Unified video annotation via multigraph learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(5):733–746, 2009. 1
- [53] Meng Wang, Hao Li, Dacheng Tao, Ke Lu, and Xindong Wu. Multimodal graph-based reranking for web image search. *IEEE transactions on image processing*, 21(11):4649–4661, 2012. 1
- [54] Xinshuo Weng and Kris Kitani. Learning spatio-temporal features with two-stream deep 3d cnns for lipreading. *arXiv preprint arXiv:1905.02540*, 2019. 3
- [55] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014. 2
- [56] Jingyun Xiao, Shuang Yang, Yuanhang Zhang, Shiguang Shan, and Xilin Chen. Deformation flow based two-stream network for lip reading. *arXiv preprint arXiv:2003.05709*, 2020. 3, 6
- [57] 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. 3
- [58] Ravindra Yadav, Ashish Sardana, Vinay P Namboodiri, and Rajesh M Hegde. Speech prediction in silent videos using variational autoencoders. *arXiv preprint arXiv:2011.07340*, 2020. 6, 7

- [59] Ryuichi Yamamoto, Martin Andrews, Michael Petrochuk, W. Hycbrom, Olga Vishnepolski, Matthew Cooper, K Chen, and Aleksas Pielikis. r9y9/wavenet vocoder: v0.1.1 release. *GitHub repository*, 2018. [7](#)
- [60] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen. Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019. [5](#), [6](#)
- [61] Yuanhang Zhang, Shuang Yang, Jingyun Xiao, Shiguang Shan, and Xilin Chen. Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition. *arXiv preprint arXiv:2003.03206*, 2020. [3](#), [6](#)
- [62] X. Zhao, S. Yang, S. Shan, and X. Chen. Mutual information maximization for effective lip reading. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG)*, pages 843–850, Los Alamitos, CA, USA, may 2020. IEEE Computer Society. [6](#)
- [63] Ya Zhao, Rui Xu, Xinchao Wang, Peng Hou, Haihong Tang, and Mingli Song. Hearing lips: Improving lip reading by distilling speech recognizers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6917–6924, 2020. [2](#)
- [64] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10394–10403, 2019. [1](#), [2](#)