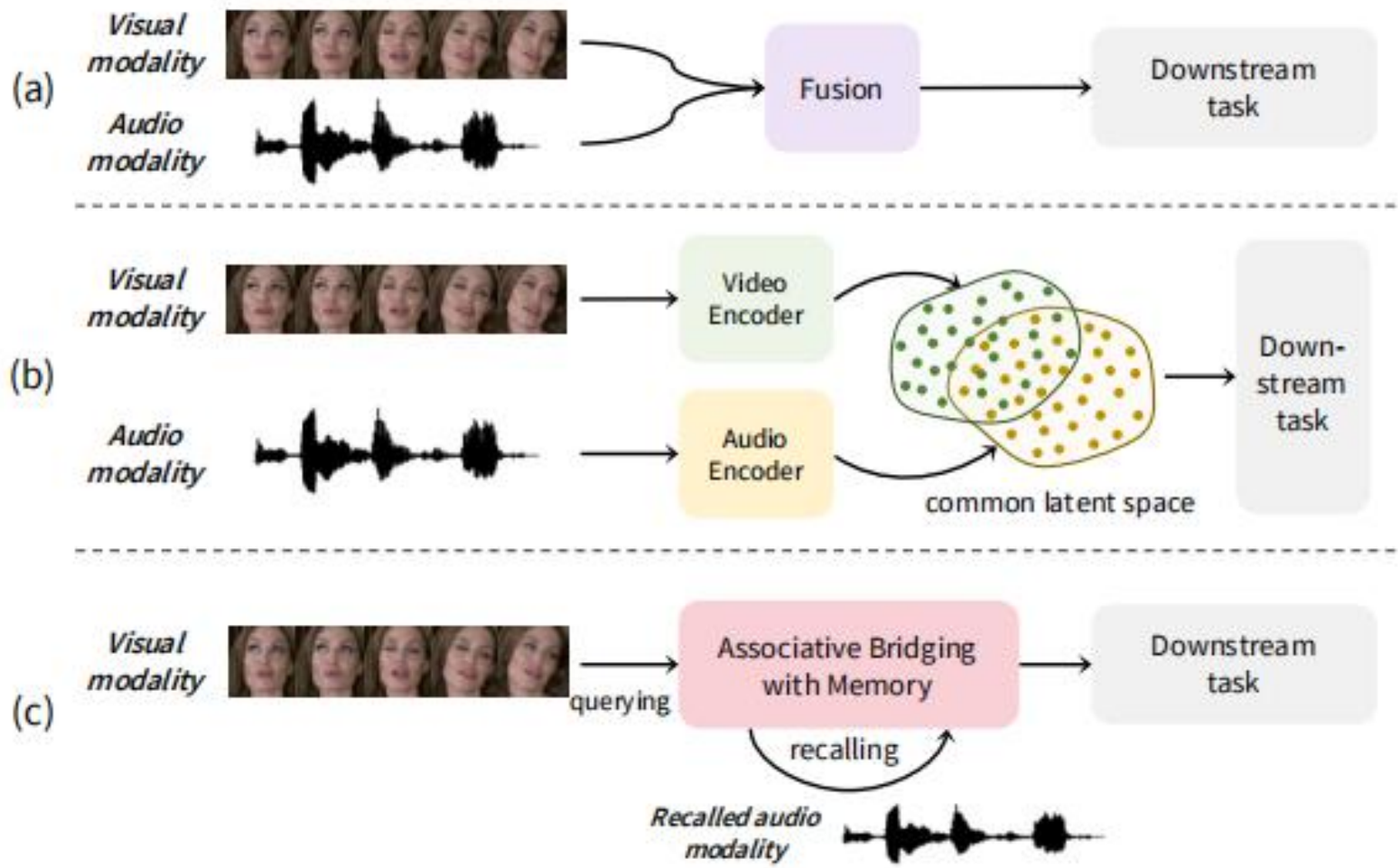
Multi-modality Associative Bridging through Memory: Speech Sound Recollected from Face Video

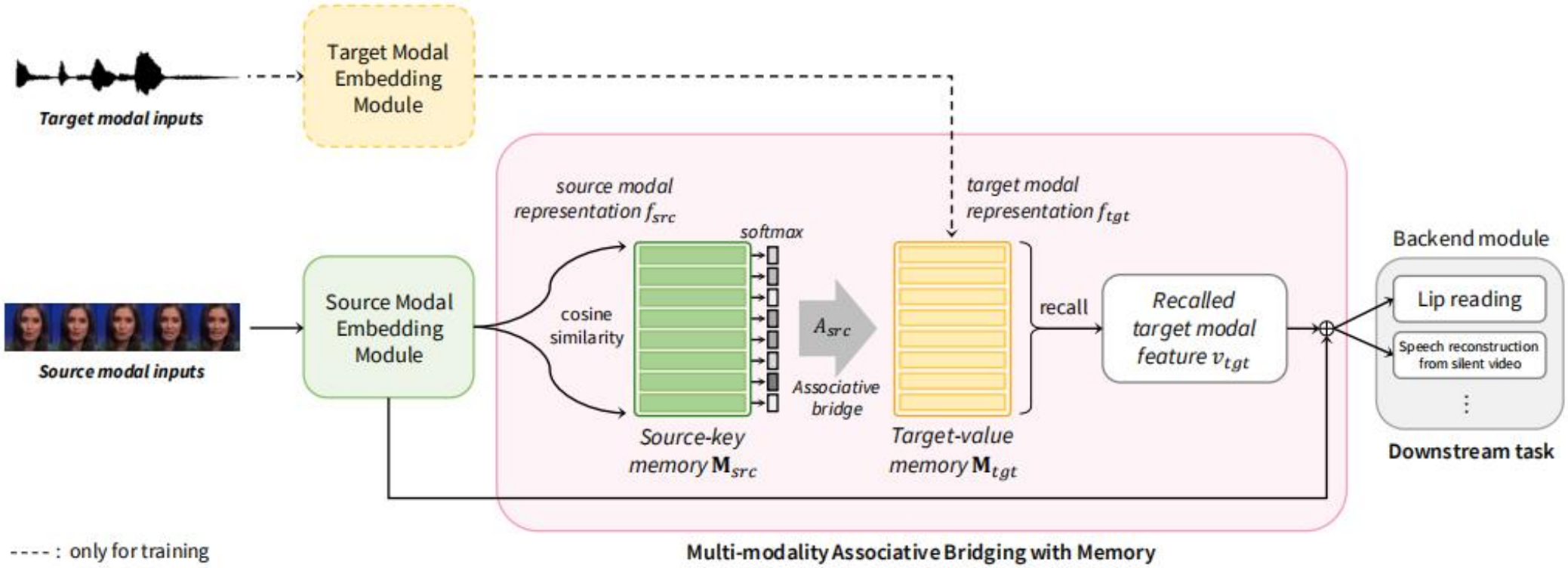
江昊宇
2022/1/19

https://openaccess.thecvf.com/content/ICCV2021/papers/
Kim_Multi-Modality_Associative_Bridging_Through_Memory_Speech_Sound_Recollected_From_Face_ICCV_2021_paper.pdf

Audio-Visual multi-modal learning



Multi-modality Associative Bridging



Multi-modality Associative Bridging

$$s_{src}^{i,j} = \frac{\mathbf{M}_{src}^{i} \cdot f_{src}^{j}}{\left\|\mathbf{M}_{src}^{i}\right\|_2 \cdot \left\|f_{src}^{j}\right\|_2} \qquad \hat{f}_{tgt}^{j} = A_{tgt}^{j} \cdot \mathbf{M}_{tgt}$$

$$\alpha_{src}^{i,j} = \frac{\exp\left(r \cdot s_{src}^{i,j}\right)}{\sum_{k=1}^{N} \exp\left(r \cdot s_{src}^{k,j}\right)} \qquad v_{tgt}^{j} = A_{src}^{j} \cdot \mathbf{M}_{tgt}$$

$$A_{src}^{j} = \left\{\alpha_{src}^{1,j}, \alpha_{src}^{2,j}, \ldots \alpha_{src}^{2,j}, \ldots \alpha_{tgt}^{N,j}\right\}$$

$$\mathcal{L}_{\text{save}} = \mathbb{E}_j\left[\left\| f_{\text{tgt}}^{j} - \hat{f}_{\text{tgt}}^{j} \right\|_2^2\right]$$

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

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

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{save}} + \mathcal{L}_{\text{bridge}} + \mathcal{L}_{\text{task}}$$

Algorithm 1 Training algorithm of the proposed framework

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.

Output: The optimized parameters of the network Φ

Randomly initialize parameters of the network Φ
for each iteration **do**
  $f_{src} = \{f_{src}^{j}\}_{j=1}^{T}$ =Source_embed(X_{src})
  $f_{tgt} = \{f_{tgt}^{j}\}_{j=1}^{T}$ =Target_embed(X_{tgt})
  for $j = 1, 2, ..., T$ **do**
    A_{src}^{j} =Softmax$(r \cdot$CosineSim$(\mathbf{M}_{src}, f_{src}^{j}))$
    A_{tgt}^{j} =Softmax$(r \cdot$CosineSim$(\mathbf{M}_{tgt}, f_{tgt}^{j}))$
    $\hat{f}_{tgt}^{j} = A_{tgt}^{j} \cdot \mathbf{M}_{tgt}$
    $v_{tgt}^{j} = A_{src}^{j} \cdot \mathbf{M}_{tgt}$
  end for
  $\mathcal{L}_{save} = \sum_{j=1}^{T} \|f_{tgt}^{j} - \hat{f}_{tgt}^{j}\|_2^2$
  $\mathcal{L}_{bridge} = \sum_{j=1}^{T} D_{KL}(A_{tgt}^{j} \| A_{src}^{j})$
  $\mathcal{L}_{task} = g(h(f_{src} \oplus v_{tgt}); y) + g(h(f_{src} \oplus f_{tgt}); y)$
  $\mathcal{L}_{tot} = \mathcal{L}_{save}/T + \mathcal{L}_{bridge}/T + \mathcal{L}_{task}$
  Update $\Phi \leftarrow \Phi - \eta \nabla_{\Phi} \mathcal{L}_{tot}$
end for

Experiments Results

- Lip reading

Method	LRW	LRW-1000
Yang *et al.* [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.

N=0, 44, 88, 132 for English and N=0, 56, 112, 168 for Mandarin.
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.

Experiments Results

- 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 *et al.* [50]	0.564	0.361	1.684
Ephrat *et al.* [12]	0.659	0.376	1.825
Lip2Wav [39]	0.731	0.535	1.772
Yadav *et al.* [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.

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

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

Moreover, with WaveNet vocoder instead of Griffin-Lim, we can improve the scores as close to that of the ground truth.

Experiments Results

- Speech reconstruction from silent video

Method	STOI	ESTOI	PESQ
Vougioukas *et al*. [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.

Experiments Results

- Learned representation inside memory

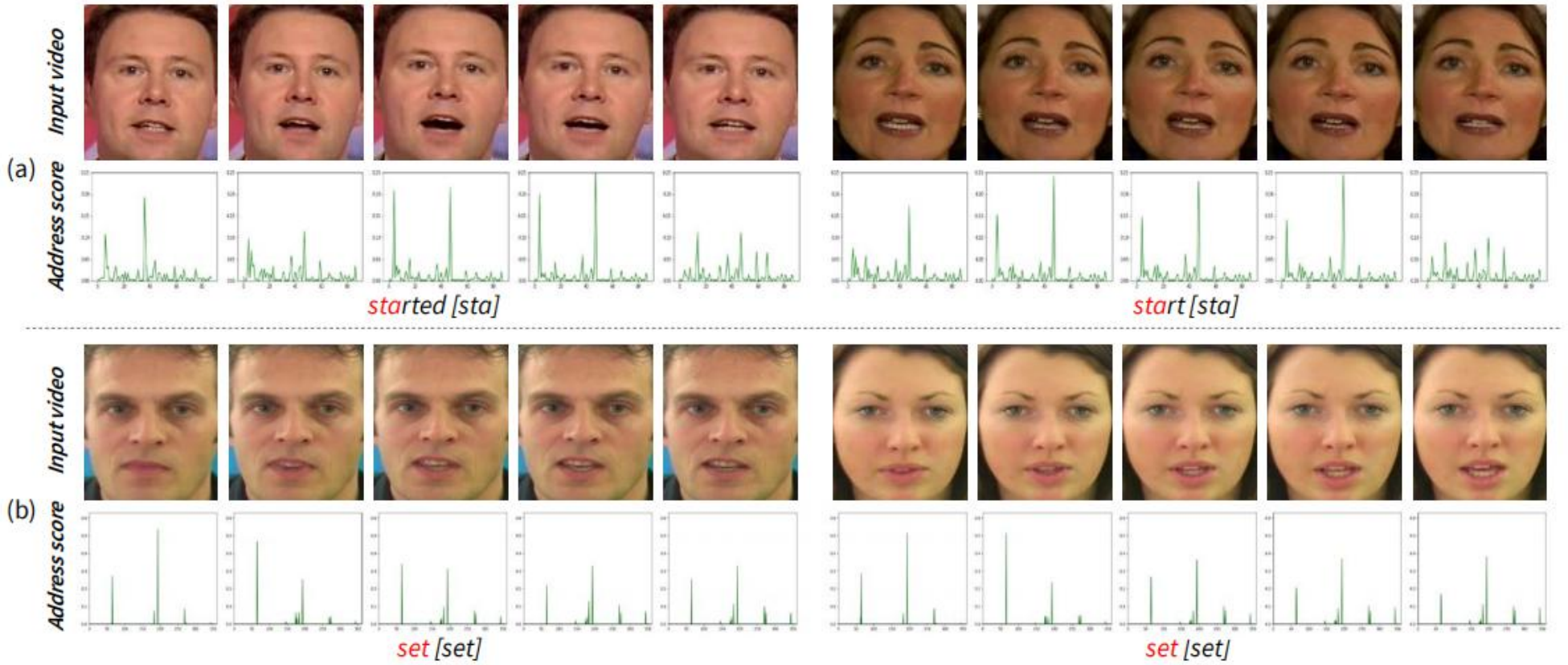


Figure 3. Face video clips (source modality) and corresponding addressing vectors for recalling audio modality (target modality) from learned representations inside memory: (a) results from lip reading and (b) results from speech reconstruction from silent video.

Experiments Results

- Learned representation inside memory

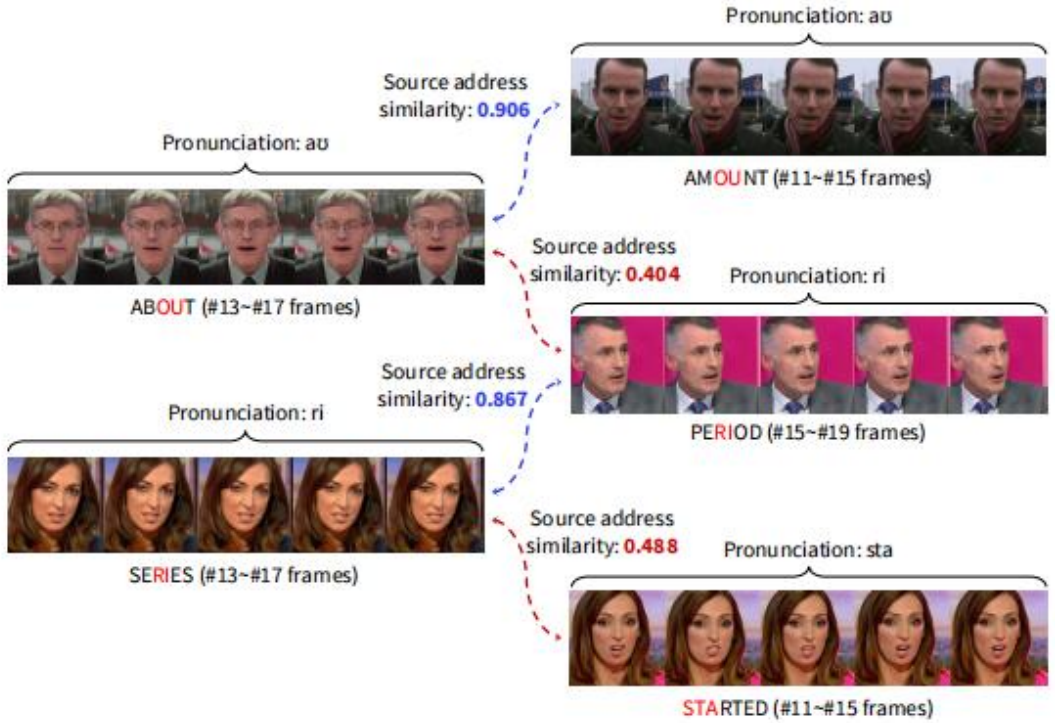


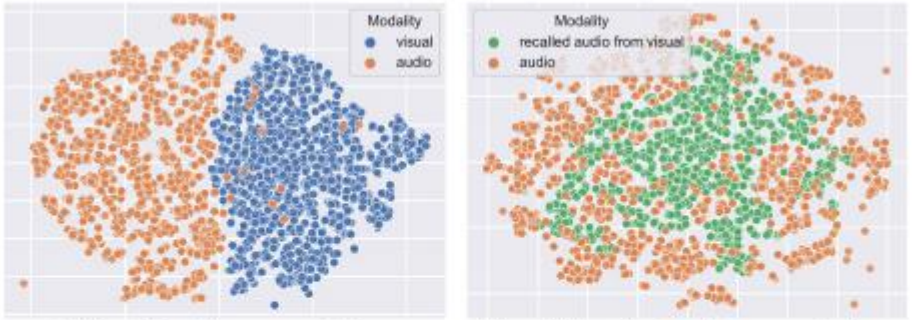
Figure 4. Examples of similarity between memory addressing vectors of different video clips in LRW. Note source addressing vector is for bridging video and audio modal features in memory.

Experiments Results

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



(a) visual & audio representations (b) recalled audio & audio representations

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.

Conclusion

- With this audio-visual multi-modal bridging framework, that can utilize both audio and visual information, even with uni-modal inputs.
- The proposed framework achieves the most advanced performance in both lip-reading and speech reconstruction from silent video.
- the proposed framework can bypass the difficulty of finding a common representation of different modalities while bridging them.

Thanks