Supplementary Material for 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

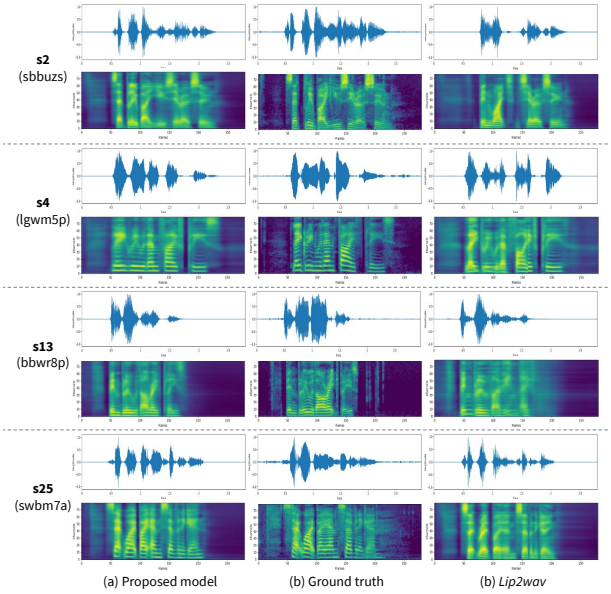

Figure 1. Qualitative comparison of GRID dataset (s2, s4, s13, s25) using (a) the proposed method, (b) the ground truth, and (c) [2] in a speaker-independent training scheme.

1. Generated results of speech reconstruction from silent video

Fig.1 shows the examples of generated mel-spectrogram from S2, S13 (male), S4, and s25 (female) in a speaker-independent setting. Since the network has not seen both the facial appearance and the lip movement, the output audio waveforms cannot be exactly the same. However, the generated mel-spectrogram and the audio waveforms seem similar to the real ones, and the actual audio sounds have shown reasonable quality. We also provide a demo for both speaker-dependent and speaker-independent settings in ***demo_dependent.mp4*** and ***demo_independent.mp4***, respectively. The demo firstly shows the silent input video. Then, the ground truth video, the generated audio from the previous work, and the generated audio from our proposed method are shown with the input video. The demo clearly indicates that the generated audio samples from the proposed model shows reasonable and correct sounds, while the previous method fails to pronounce the perfectly correct letters. We write in red on letters with the wrong sounds in the actual transcription at the bottom of the demo video screen. Furthermore, our generated audio samples clearly follow the voice with the correct gender corresponding to the input face video, while the previous work [2] often fails (*e.g.*, s25 in ***demo_independent.mp4***).

2. Visualization of addressing vectors for additional talking input video

We visualize the addressing vectors of both lip reading model and speech reconstruction model in a speaker-independent setting, additional to the *Fig.3* in the manuscript. Fig.2(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 that the tendency of the addressing vectors is similar. The same tendency can be observed in the speech reconstruction model which is shown in Fig.2(b). In addition, we make a demo video showing changes in key-value memory addressing during training on LRW dataset. The demo video ***demo_address.mp4*** clearly shows that the source-key memory addressing vector well follows the target-value memory addressing vector as epoch increases. It also indicates that the tendencies of both the addressing vectors of two similar pronouncing videos are similar to each other.

3. Network architecture

Application 1. Lip reading: The architectures of both visual embedding module and audio embedding module are shown on Table 1. The audio embedding module is composed of two convolution layers with stride 2, one Residual block [1] with stride 1, and a fully connected layer, which aggregates the spectral dimension and channel di-

*Both authors have contributed equally to this work.
†Corresponding author

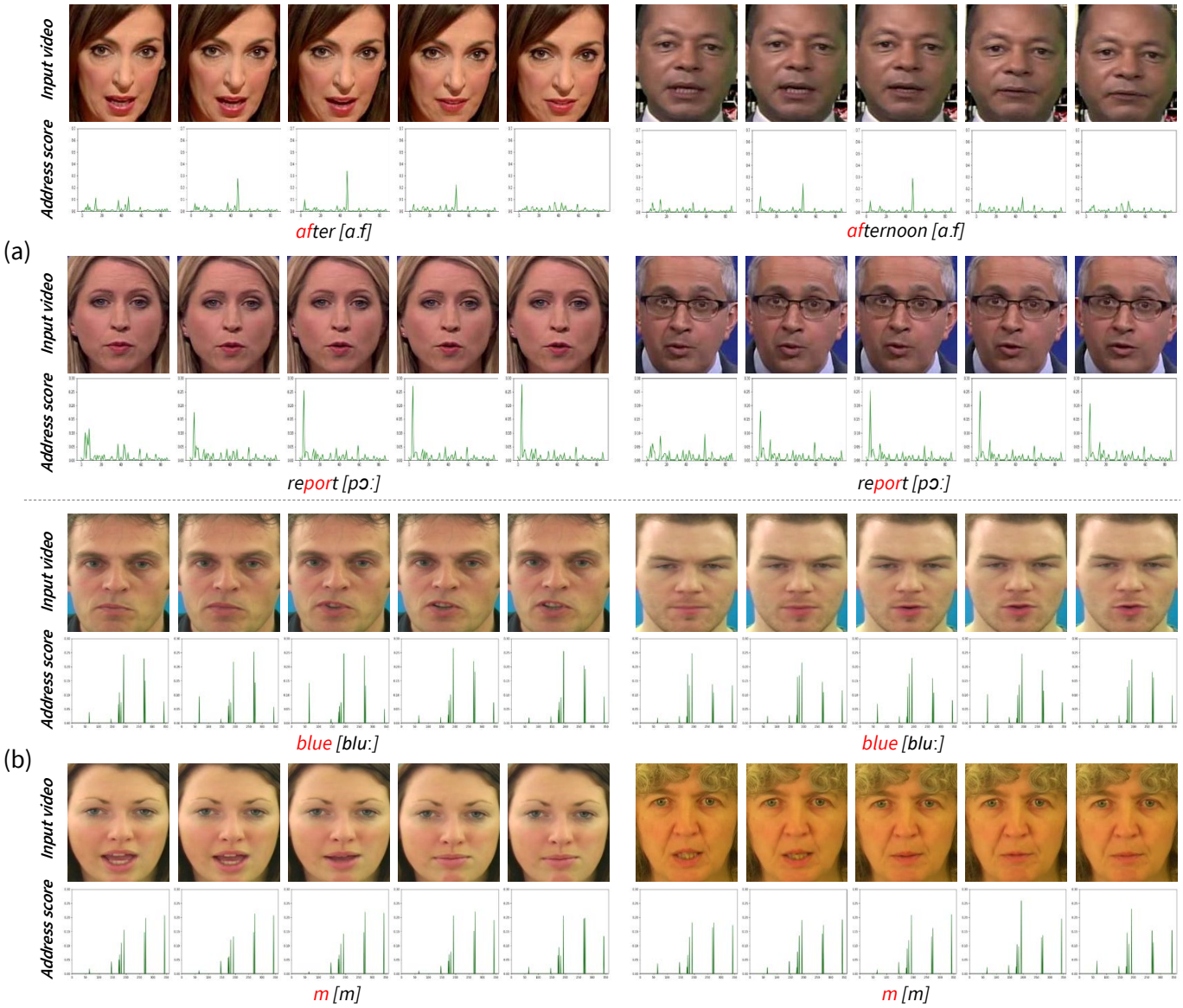

Figure 2. Face video clips (source modality) and corresponding addressing vectors for audio modality (target modality) from learned representations inside memory: (a) results from lip reading and (b) results from speech reconstruction from silent video.

mension. The dimension of both embedded representations are 512 (*i.e.*, $C = D = 512$). For the fusion layer $h(\cdot)$, we use one linear layer, and 16 is used for r. Since the baseline visual embedding module encodes a short-time range (*i.e.*, 5 frames) with one 3D convolution layer, we examine the number of memory slots using multiples of the number of pronunciations for each language as a hint (*i.e.*, 44 phonemes for English and 56 pinyins for Mandarin). We find that the doubles of the number of pronunciations achieve the best among the variants for both languages (*i.e.*, $N = 88$ for English, $N = 112$ for Mandarin).

Application 2. Speech reconstruction from silent video: The architectures of both visual embedding module and audio embedding module for speech reconstruction task are shown on Table 2. 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. The dimension of both embedded representations are 512 (*i.e.*, $C = D = 512$). For the fusion layer $h(\cdot)$, we use one linear layer, and 16 is used for r. Since the baseline visual embedding module encodes a long-time range with stacked 3D convolution layers, we use larger number of memory slot size than lip reading experiment (*i.e.*, $N = 150$).

Visual Embedding Module: input size T × H × W × 1		
Layer	**Filter size / number / stride**	**Output dimensions**
Conv 3D	$5 \times 7 \times 7$ / 64 / [1, 2, 2]	$T \times \frac{H}{2} \times \frac{W}{2} \times 64$
Max Pool 3D	$1 \times 3 \times 3$ / - / [1, 2, 2]	$T \times \frac{H}{4} \times \frac{W}{4} \times 64$
ResBlock 2D	3×3 / 64 / [1, 1] 3×3 / 64 / [1, 1]	$T \times \frac{H}{4} \times \frac{W}{4} \times 64$
ResBlock 2D	3×3 / 64 / [1, 1] 3×3 / 64 / [1, 1]	$T \times \frac{H}{4} \times \frac{W}{4} \times 64$
ResBlock 2D	3×3 / 128 / [2, 2] 3×3 / 128 / [1, 1]	$T \times \frac{H}{8} \times \frac{W}{8} \times 128$
ResBlock 2D	3×3 / 128 / [1, 1] 3×3 / 128 / [1, 1]	$T \times \frac{H}{8} \times \frac{W}{8} \times 128$
ResBlock 2D	3×3 / 256 / [2, 2] 3×3 / 256 / [1, 1]	$T \times \frac{H}{16} \times \frac{W}{16} \times 256$
ResBlock 2D	3×3 / 256 / [1, 1] 3×3 / 256 / [1, 1]	$T \times \frac{H}{16} \times \frac{W}{16} \times 256$
ResBlock 2D	3×3 / 512 / [2, 2] 3×3 / 512 / [1, 1]	$T \times \frac{H}{32} \times \frac{W}{32} \times 512$
ResBlock 2D	3×3 / 512 / [1, 1] 3×3 / 512 / [1, 1]	$T \times \frac{H}{32} \times \frac{W}{32} \times 512$
Avg Pool 2D	$\frac{H}{32} \times \frac{W}{32}$ / - / [1, 1]	T × 512

Audio Embedding Module: input size M × 4T × 1		
Layer	**Filter size / number / stride**	**Output dimensions**
Conv 2D	3×3 / 128 / [2, 2]	$\frac{M}{2} \times 2T \times 128$
Conv 2D	3×3 / 256 / [2, 2]	$\frac{M}{4} \times T \times 256$
ResBlock 2D	3×3 / 256 / [1, 1] 3×3 / 256 / [1, 1]	$\frac{M}{4} \times T \times 256$
Flatten	-	$T \times \frac{M}{4} * 256$
Linear	$\frac{M}{4} * 256 \times 512$	T × 512

Table 1. Network architecture for lip reading.

Visual Embedding Module: input size T × H × W × 3		
Layer	**Filter size / number / stride**	**Output dimensions**
Conv 3D	$5 \times 5 \times 5$ / 32 / [1, 2, 2]	$T \times \frac{H}{2} \times \frac{W}{2} \times 32$
Conv 3D	$3 \times 3 \times 3$ / 32 / [1, 1, 1]	$T \times \frac{H}{2} \times \frac{W}{2} \times 32$
Conv 3D	$3 \times 3 \times 3$ / 32 / [1, 1, 1]	$T \times \frac{H}{2} \times \frac{W}{2} \times 32$
Conv 3D	$3 \times 3 \times 3$ / 64 / [1, 2, 2]	$T \times \frac{H}{4} \times \frac{W}{4} \times 64$
Conv 3D	$3 \times 3 \times 3$ / 64 / [1, 1, 1]	$T \times \frac{H}{4} \times \frac{W}{4} \times 64$
Conv 3D	$3 \times 3 \times 3$ / 64 / [1, 1, 1]	$T \times \frac{H}{4} \times \frac{W}{4} \times 64$
Conv 3D	$3 \times 3 \times 3$ / 128 / [1, 2, 2]	$T \times \frac{H}{8} \times \frac{W}{8} \times 128$
Conv 3D	$3 \times 3 \times 3$ / 128 / [1, 1, 1]	$T \times \frac{H}{8} \times \frac{W}{8} \times 128$
Conv 3D	$3 \times 3 \times 3$ / 128 / [1, 1, 1]	$T \times \frac{H}{8} \times \frac{W}{8} \times 128$
Conv 3D	$3 \times 3 \times 3$ / 256 / [1, 2, 2]	$T \times \frac{H}{16} \times \frac{W}{16} \times 256$
Conv 3D	$3 \times 3 \times 3$ / 256 / [1, 1, 1]	$T \times \frac{H}{16} \times \frac{W}{16} \times 256$
Conv 3D	$3 \times 3 \times 3$ / 256 / [1, 1, 1]	$T \times \frac{H}{16} \times \frac{W}{16} \times 256$
Conv 3D	$3 \times 3 \times 3$ / 512 / [1, 2, 2]	$T \times \frac{H}{32} \times \frac{W}{32} \times 512$
Conv 3D	$3 \times 3 \times 3$ / 512 / [1, 1, 1]	$T \times \frac{H}{32} \times \frac{W}{32} \times 512$
Conv 3D	$3 \times 3 \times 3$ / 512 / [1, 1, 1]	$T \times \frac{H}{32} \times \frac{W}{32} \times 512$
Conv 3D	$3 \times 3 \times 3$ / 512 / [1, 1, 1]	$T \times \frac{H}{32} - 2 \times \frac{W}{32} - 2 \times 512$
BiLSTM	-	T × 512

Audio Embedding Module: input size M × 4T × 1		
Layer	**Filter size / number / stride**	**Output dimensions**
Conv 2D	3×3 / 128 / [2, 2]	$\frac{M}{2} \times 2T \times 128$
Conv 2D	3×3 / 256 / [2, 2]	$\frac{M}{4} \times T \times 256$
Conv 2D	5×5 / 256 / [1, 1]	$\frac{M}{4} \times T \times 256$
ResBlock 2D	3×3 / 256 / [1, 1] 3×3 / 256 / [1, 1]	$\frac{M}{4} \times T \times 256$
Flatten	-	$T \times \frac{M}{4} * 256$
Linear	$\frac{M}{4} * 256 \times 512$	T × 512

Table 2. Network architecture for speech reconstruction from silent video.

Dataset	N=0	N=44	N=88	N=132
LRW	84.14%	85.34%	**85.41%**	85.32%
Dataset	N=0	N=56	N=112	N=168
LRW-1000	44.93%	50.70%	**50.82%**	50.70%

Table 3. Lip reading: ablation on memory slot size

Metric	N=0 [2]	N=50	N=150
STOI	0.731	0.731	**0.738**
ESTOI	0.535	0.570	**0.579**
PESQ	1.772	1.946	**1.984**

Table 4. Speech reconstruction from silent video: ablation on memory slot size

4. Ablation study on memory slot size

Application 1. Lip reading: In order to examine the effect of the number of memory slots, we conduct an ablation study with four different number of memory slots for each dataset. The ablation results on memory slot size are reported at Table 3. For LRW, a dataset in English, the word accuracy is achieved the best with 85.41% when N=88. It increases the performance from the baseline without the proposed framework by 1.27%. For LRW-1000, a dataset in Mandarin, 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. Moreover, we observed that the double number of phoneme and pinyin is the best performance for each language (*i.e.*, 88 for English, 112 for Mandarin).

Application 2. Speech reconstruction from silent video: Table 4 shows the ablation results on differentiating memory slot sizes (*i.e.*, 0, 50, 150) in the speaker-dependent setting. As the table shows, the performances in three qualitative metrics are improved disregarding the number of memory slots, which verifies the effectiveness of the proposed method, and we report $N = 150$ for the experiment in speaker-dependent setting on GRID. This indicates that the larger memory slot size becomes, the greater the model performance achieves. Thus, we use $N = 360$ for the speaker-independent setting which has more variables for different lip movements and the corresponding speech sounds.

References

[1] 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 recog-*

nition, pages 770–778, 2016. 1

[2] 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