Supplementary: Lip to Speech Synthesis with Visual Context Attentional GAN

1 Architectural Details

In this section, we describe the detailed architecture of each module in the proposed method. The architectures of local visual encoder (ϕ_v), global visual encoder (ϕ_c), generators (ψ_1, ψ_2, ψ_3), discriminators (D_1, D_2, D_3), local audio encoder (ϕ_a), and postnet are described in Table 1, 2, 3, 4, 5, and 6, respectively. For the ResBlock, we denote the first convolution layer only, and the second convolution layer is omitted (It has the same filter size and number with stride 1). The stride 2 in ResBlock of the generators indicates upsample; otherwise, it represents downsample. Moreover, the output size of speech representation is represented with 80 mel-spectral dimension (*i.e.*, F=80). For the weights in audio-visual attention, $W_q^i \in \mathbb{R}^{2560\times128}$, $W_k^i \in \mathbb{R}^{512\times128}$, and $W_v^i \in \mathbb{R}^{512\times1280}$ are utilized for two audio-visual attentions. Note that the local visual encoder and the local audio encoder are designed to share the same temporal receptive fields (*i.e.*, about 0.2-sec for 25fps video).

2 Visualization of attention map

Fig. 1 shows visualization of the attention map of the second audio-visual attention in VCA-GAN. The x-axis of the attention map represents generating audio frames and the y-axis represents video frames. We also visualize the corresponding generated mel-spectrogram at the second generator. Note that the attended visual frames are observed in accordance with the corresponding audio frames (two times frame number of video frame number). Moreover, the attended frames cover a somewhat wide range which means the audio-visual attention tends to attend to global visual features in neighbors of the current audio frame.

3 Used pre-trained ASR model

The pre-trained ASR model used for evaluating WER is based on [55] and modified into lighter architecture, shown in Table 7. It is trained on each setting on each dataset (*i.e.*, constrained-speaker setting of GRID, unseen-speaker setting of GRID, multi-speaker setting of GRID, and LRW) using CTC loss [24] for GRID and cross-entropy loss for LRW. Each model has WER on its test set, 0.83%, 1.67%, 0.37%, 1.54%, respectively, with the same order in the above bracket. All the pre-trained ASR models can be found in the supplemental material and are publicly available for fair comparison.

4 Qualitative results of ablation study

Fig. 2 shows the qualitative results of ablation study. The dotted red-line is to check how well the generated mel-spectrogram is in-sync with the ground-truth mel-spectrogram. It is clearly indicated that the modules including the synchronization technique (*i.e.*, + synchronization, + multi-discriminators) are much in sync with the ground-truth mel-spectrogram. Further, the red box of the figures on the right proves that the more modules are included, the more detailed mel-spectrogram is generated. This confirms the effectiveness of each proposed module qualitatively.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

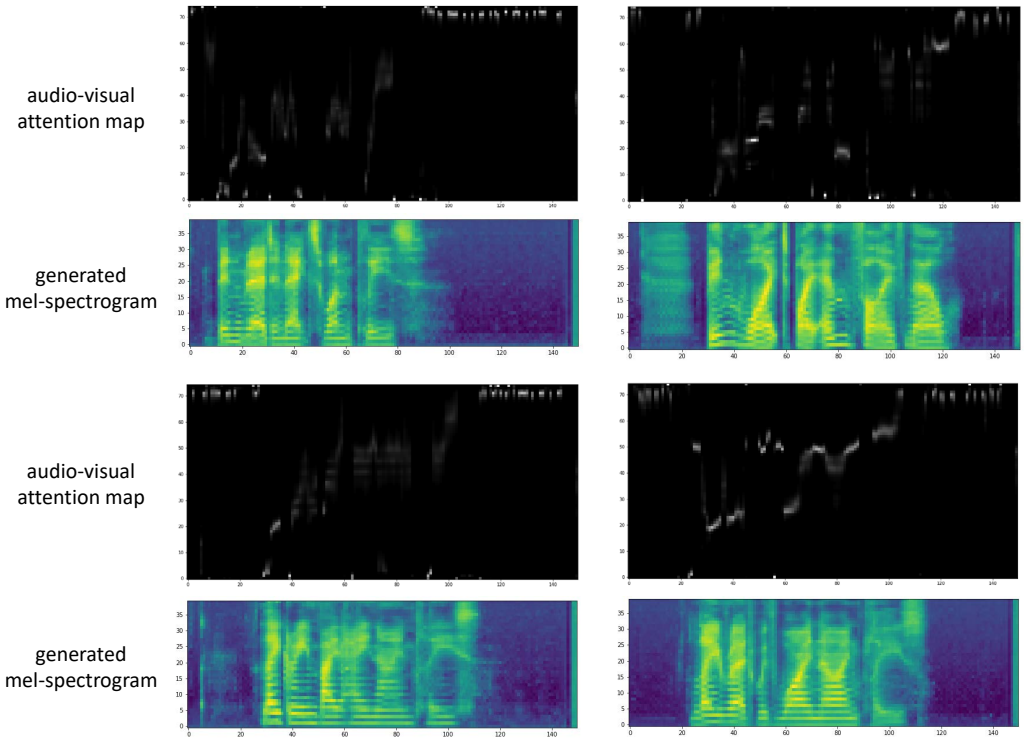

Figure 1: Attention map of audio-visual attention and corresponding generated mel-spectrogram at 2nd generator.

5 Additional results

We additionally visualize the qualitative results of 1D GAN-based [10], Lip2Wav [4], Vocoder-based [12], and the proposed VCA-GAN in Fig. 3 and 4. Moreover, the demo video is available in the supplemental material.

Table 1: Architecture of Local Visual Encoder

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

Table 2: Architecture of Global Visual Encoder

Global Visual Encoder: input size T × D		
Layer	**Hidden dim**	**Output dimensions**
Bi-GRU	512	T × 1024
Bi-GRU	512	T × 1024
Linear	512	T × 512

Table 3: Architecture of Generator

Generator: input size 20 × T × (D+128)			
model	**Layer**	**Filter size / number / stride**	**Output dimensions**
ψ_1	ResBlock 2D	5 × 5 / 512 / [1, 1]	20 × T × 512
	ResBlock 2D	5 × 5 / 256 / [1, 1]	20 × T × 256
	ResBlock 2D	5 × 5 / 256 / [1, 1]	20 × T × 256
	ResBlock 2D	5 × 5 / 128 / [1, 1]	20 × T × 128
	ResBlock 2D	5 × 5 / 128 / [1, 1]	20 × T × 128
	ResBlock 2D	5 × 5 / 128 / [1, 1]	20 × T × 128
ψ_2	visual context attention	-	20 × T × (128 + 64)
	Conv 2D	5 × 5 / 128 / [1, 1]	20 × T × 128
	ResBlock 2D	5 × 5 / 64 / [2, 2]	40 × 2T × 64
	ResBlock 2D	5 × 5 / 64 / [1, 1]	40 × 2T × 64
	ResBlock 2D	5 × 5 / 64 / [1, 1]	40 × 2T × 64
ψ_3	visual context attention	-	40 × 2T × (64 + 32)
	Conv 2D	5 × 5 / 64 / [1, 1]	40 × 2T × 64
	ResBlock 2D	5 × 5 / 32 / [2, 2]	80 × 4T × 32
	ResBlock 2D	5 × 5 / 32 / [1, 1]	80 × 4T × 32
	ResBlock 2D	5 × 5 / 32 / [1, 1]	80 × 4T × 32
To mel ψ_1, ψ_2, ψ_3	Conv 2D	1 × 1 / 1 / [1, 1]	80 × 4T × 1

Table 4: Architecture of Discriminator

Discriminator: input size of D_3 80 × T × 1			
model	**Layer**	**Filter size / number / stride**	**Output size of D_3**
D_1, D_2, D_3	ResBlock 2D	5 × 5 / 32 / [2, 2]	40 × $\frac{T}{2}$ × 32
	ResBlock 2D	5 × 5 / 64 / [2, 2]	20 × $\frac{T}{4}$ × 64
D_2, D_3	ResBlock 2D	5 × 5 / 128 / [2, 2]	10 × $\frac{T}{8}$ × 128
D_3	ResBlock 2D	5 × 5 / 256 / [2, 2]	5 × $\frac{T}{16}$ × 256
unconditioned $\{D_1, D_2, D_3\}$	Conv 2D	5 × 5 / {64, 128, 256} / [1, 1]	1 × $\frac{T}{16}$-4 × 256
	Avg Pool 2D	-	256
	Linear	1	1
conditioned $\{D_1, D_2, D_3\}$	Cat w/ $\mathcal{M}(C_v)$	-	5 × $\frac{T}{16}$ × (256 + 512)
	Conv 2D	5 × 5 / {64, 128, 256} / [1, 1]	5 × $\frac{T}{16}$ × 256
	Conv 2D	5 × 5 / {64, 128, 256} / [1, 1]	1 × $\frac{T}{16}$-4 × 256
	Avg Pool 2D	-	256
	Linear	1	1

Table 5: Architecture of Local Audio Encoder

Local Audio Encoder: input size 80 × 4T × 1		
Layer	**Filter size / number / stride**	**Output dimensions**
Conv 2D	3 × 3 / 128 / [2, 2]	40 × 2T × 128
Conv 2D	3 × 3 / 256 / [2, 2]	20 × T × 256
ResBlock 2D	3 × 3 / 256 / [1, 1]	20 × T × 256
Flatten	-	T × 20 * 256
Linear	512	T × 512

Table 6: Architecture of Postnet

Postnet: input size 80 × 4T, $\mathbb{F}$: size of FFT		
Layer	**Filter size / number / stride**	**Output dimensions**
Conv 1D	7 / 128 / 1	128 × 4T
ResBlock 1D	5 / 256 / 1	256 × 4T
ResBlock 1D	5 / 256 / 1	256 × 4T
ResBlock 1D	5 / 256 / 1	256 × 4T
Conv 1D	1 / $\mathbb{F}$ / 1	$\mathbb{F}$ × 4T

Table 7: Architecture of ASR model

ASR: input size 80 × 4T × 1		
Layer	**Filter size / number / stride**	**Output dimensions**
Conv 2D	3 × 3 / 32 / [2, 2]	40 × 2T × 32
Conv 2D	3 × 3 / 64 / [2, 2]	20 × T × 64
ResBlock 2D	3 × 3 / 64 / [1, 1]	20 × T × 64
Flatten	-	T × 20 * 64
Linear	256	T × 256
Bi-GRU	256	T × 512
Bi-GRU	256	T × 512
Linear	512	T × class_num

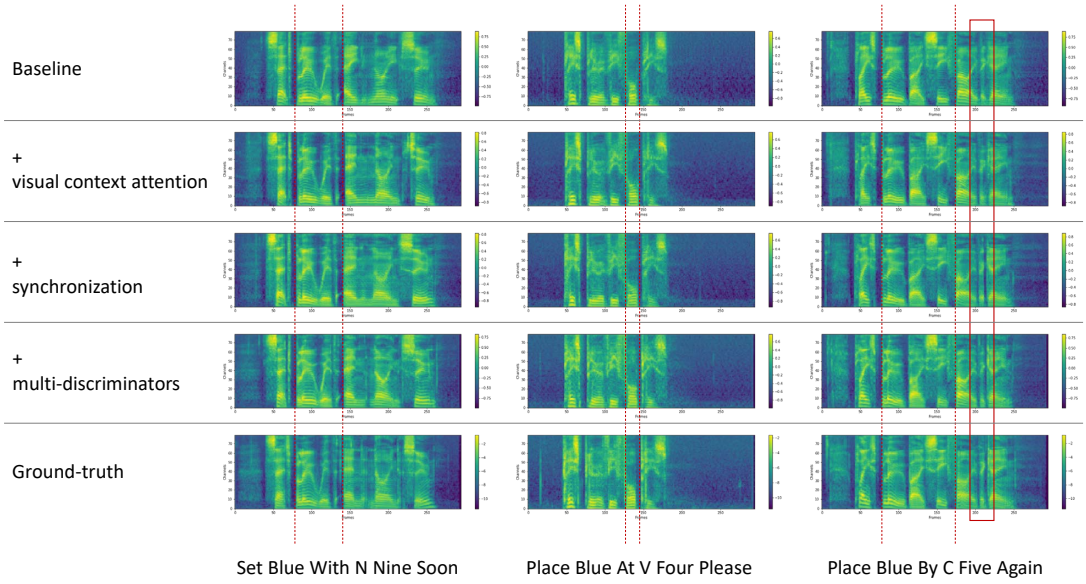

Figure 2: Mel-spectrogram comparison results of the ablation study.

6 Instructions used for MOS experiment

Fig. 5 shows the screenshot of the instructions used for MOS experiment. 12 volunteers were recruited online, and the evaluation was performed through Google Doc. A total of 300-sec videos were provided, and the expected quality assessment time was 10 minutes per person.

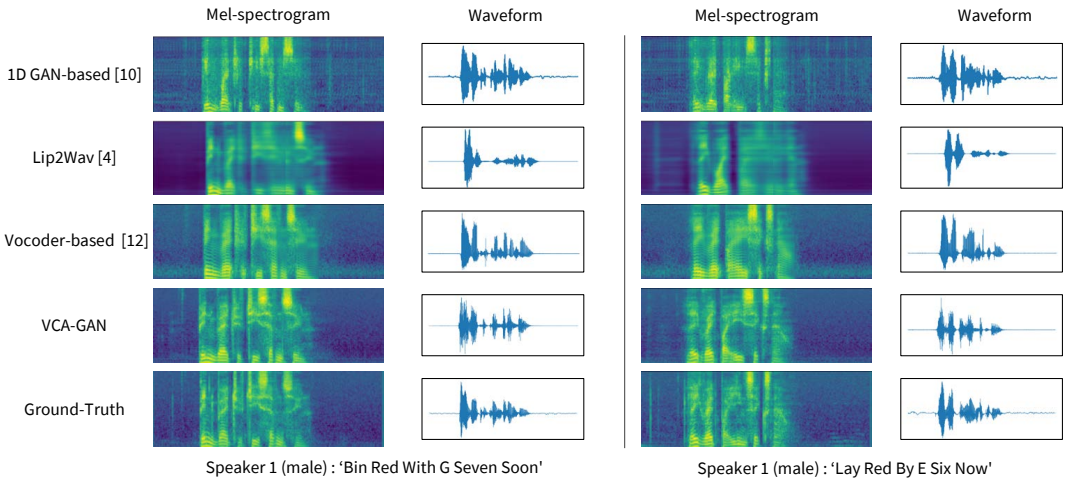

Figure 3: Additional qualitative results.

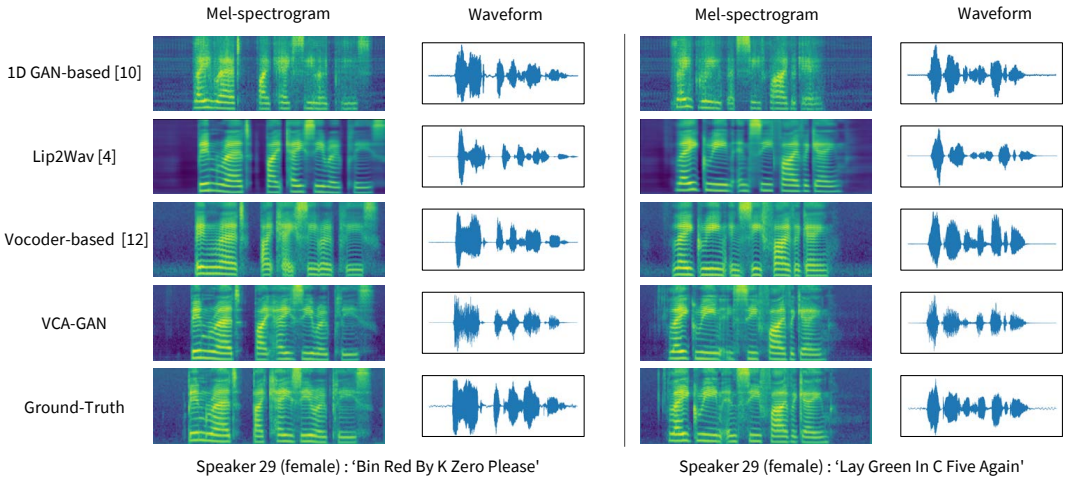

Figure 4: Additional qualitative results.

Speech synthesis quality human evaluation

This survey will be used for research purposes, evaluating naturalness, intelligibility, and sync matching of the provided speech.

A total of 100 utterances and corresponding sentences are provided.
The sentence consists of six word combinations in order (command->color->preposition->letter->digit->adverb), as shown in the table below. (e.g., place blue at A 8 now/ lay blue in A 8 soon)

We would appreciate it if you could listen to the given voice for each question and answer the questions in 1~5 score while comparing with the sentences.

1) Naturalness: whether speech is similar to natural human speech, hence items such as naturalness, ease of
listening, pleasantness and audio flow are relevant
2) Intelligibility: whether words and sentences can be
understood, therefore, items tapping into the factor assess listening effort, pronunciation, speaking rate, comprehension problems, and articulation
3) Sync Matching:
whether words and lip movements are in sync

We are grateful for having you participating our survey.

<GRID sentence grammar>

Command	Color	Preposition	Letter	Digit	Adverb
bin	blue	at	A-Z	0-9	again
lay	green	by	minus W		now
place	red	in			please
set	white	with			soon

Figure 5: Instructions used for MOS.