VisageSynTalk: Unseen Speaker Video-to-Speech Synthesis via Speech-Visage Feature Selection

Supplementary Material

Table 1. Performance comparison in audio-visual synchronization metrics, LSE-D (lower is better) and LSE-C (higher is better) in multi-speaker independent setting on GRID corpus

Method	LSE-D(↓)	LSE-C(↑)
GAN-based [13]	8.318	4.607
Vocoder-based [8]	7.884	4.349
Lip2Wav [10]	7.103	6.441
VV-Memory [6]	7.227	6.146
End-to-end GAN [9]	7.006	6.112
Proposed model	**6.955**	**6.475**

Table 2. Performance comparison in 4-speaker (s1, s2, s4, s29) dependent setting on GRID corpus

Method	STOI	ESTOI	PESQ
GAN-based [13]	0.565	0.318	1.483
Vocoder-based [8]	0.648	0.447	1.626
Lip2Wav [10]	0.649	0.469	1.678
VV-Memory [6]	0.652	0.476	1.792
End-to-end GAN [9]	0.647	0.436	1.691
Proposed model	**0.668**	**0.541**	**1.855**

1 Additional quantitative results

1.1 Speech content recognition performance

We verify the speech content recognition performance using audio-visual synchronization rate of the generated speech with the input lip sequences. We adopt Lip Sync Error-Distance (LSE-D) and Lip Sync Error-Coinfidence(LSE-C) metrics from [11]. LSE-D denotes a matching distance between audio and video that the lower LSE-D corresponds to the higher audio-visual match, i.e., the speech and lip movements are in sync, and LSE-C refers to the average confidence score that the lower confidence score denotes that there are several completely out-of-sync parts. Table 1 shows the performance of the results on multi-speaker independent setting in GRID corpus dataset. The proposed framework outperforms the previous work, attaining 6.955 LSE-D and 6.475 LSE-C.

Table 3. Analysis on different number of speech selective masks in multi-speaker independent setting on GRID dataset

Metric	N=1	N=3	**N=6**	N=9
STOI	0.556	0.540	**0.567**	0.561
ESTOI	0.291	0.304	**0.308**	0.260
PESQ	1.360	1.368	**1.373**	1.372

Table 4. Performance comparison in multi-speaker independent setting on LRW

Method	STOI	ESTOI	PESQ
Lip2Wav (w/o speaker embedding) [10]	0.375	0.026	1.198
Lip2Wav (w/ speaker embedding) [10]	0.511	0.301	1.260
VV-Memory [6]	0.548	0.264	1.256
Proposed model	**0.555**	**0.305**	**1.264**

1.2 Results in GRID corpus: 4-speaker dependent setting

We additionally conduct the experiment on 4-speaker setting of the GRID dataset, where subject 1, 2, 4, and 29 are trained together. This setting is generally utilized in the early studies [7–9, 13]. Table 2 indicates the comparison results with the previous methods [7–10, 13]. The proposed framework outperforms all the early works by achieving 0.668, 0.541, and 1.855, on STOI, ESTOI, and PESQ, respectively.

1.3 Results in LRW: multi-speaker independent setting

We compare the performacne of our model in multi-speaker independent (unseen) setting of LRW dataset with the previous works [6, 10]. Since Lip2Wav [11] utilizes additional speaker information and feed a speaker embeddings input to the model in both training and testing, we conduct experiment using Lip2Wav model with and without speaker embedding information for fair comparison. Clearly, Lip2Wav does not perform well without speaker embedding. Table 4 verifies that our model outperforms all previous works [6, 10], showing even better performance than Lip2Wav with speaker embedding.

1.4 Effectiveness of multi-heads

Furthermore, we check the performance by differing the number of heads in the multi-speaker independent (unseen) setting on GRID corpus dataset. As shown in Table 3, our model achieves 0.556 STOI, 0.291 ESTOI, and 1.360 PESQ when using the single speech selective mask. It attains the best performance when utilizing 6 speech selective mask, obtaining the score of 0.567, 0.308, and 1.373 on STOI, ESTOI, and PESQ, respectively. When the number of masks increases to 9, the performances slightly decrease to 0.561 STOI, 0.260 ESTOI, and 1.372 PESQ, meaning that the performances saturates when the sufficient number of speech selective masks are provided. The proposed model achieves the best performances with 9 speech selective masks on the multi-speaker dependent setting (Table 6 in the manuscript). Compared to the performances on the multi-speaker dependent setting, we can assume that the number of speech selective

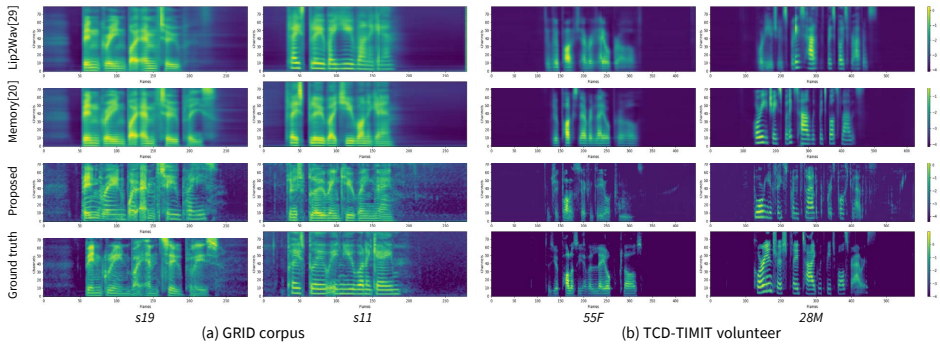

Fig. 1. Additional visualization examples of generated mel-spectrogram from GRID corpus and TCD-TIMIT volunteer datasets in speaker-independent settings

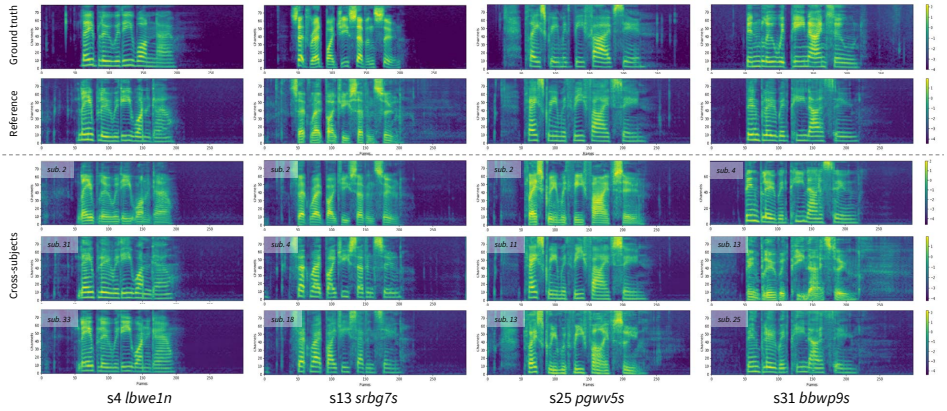

Fig. 2. Additional qualitative results of the ground truth and the generated mel-spectrogram by changing the reference speaking-style features of subject ids

masks depend on the number of speakers, since the multi-speaker dependent setting uses total 33 speakers in training where the independent setting utilizes much less number of speakers. This means that 6 speech selective masks are enough to contain diverse characteristics of different subjects for multi-speaker independent (unseen) setting.

2 Additional qualitative results

2.1 Results in multi-speaker independent setting

Fig. 1 shows the additional visualization examples of generated mel-spectrogram from GRID corpus and TCD-TIMIT volunteer datasets in multi-speaker independent settings where unseen speakers are taken in the inference time. We also provide a GRID demo for both multi-speaker independent and multi-speaker dependent settings in ***demo_independent.mp4*** and ***demo_dependent.mp4***, respectively, and LRW demo in ***demo_independent_lrw.mp4***. The demo firstly shows the silent input video. Then, the ground truth video, the generated audio

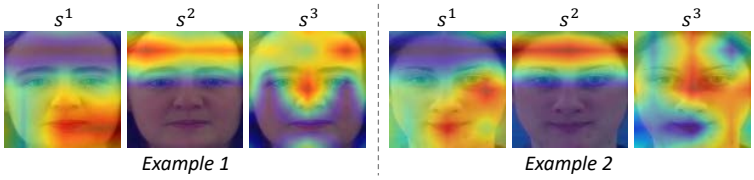

Fig. 3. Examples of visualization of attribution maps using GRID corpus test dataset. Red is the most attributed region, and blue is the least attributed region.

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 and do not match voices. We write in red on letters with the wrong sounds in the actual transcription at the bottom of the demo video screen.

2.2 Results altering the visage-styles

We demonstrate extra examples of the synthesized mel-spectrogram by changing the visage-style features of subject ids in multi-speaker independent setting. Fig. 2 shows the results of the generated mel-spectrograms with 3 different visage-style features, which are originally from subject id 4, 13, 25, 31 of the GRID corpus dataset, respectively. We further provide a demo for cross-speaker setting in ***demo_altering_visage_styles.mp4***. The demo firstly shows the ground truth video. Then, the simplified scheme of generating speech from the actual subejct's speech content and visage-style is shown, and the generated audio samples from the actual subject's speech content $\boldsymbol{f}_{sc}$ and the visage-styles $\boldsymbol{s}^*$'s from 3 different subjects are demonstrated sequentially. The results clearly show that when adopting the visage-style from other subject identity, the audio speech well mimics voice of the input visage style while maintaining the content of the speech. This verifies that the proposed speech-visage feature selection module can separate the speech content and the identity from the input silent video, and the VS-synthesizer is able to reconstruct the speech given the speech content and the visage style features.

2.3 Visualization of each visage-style feature

To explore each visage-style feature, we utilize Grad-CAM [12] and visualize attribution maps by activating each s^1, s^2, and s^3 independently while suppressing others. Using GRID test dataset of multi-speaker independent setting, we attain attribution maps of the last convolution layer in the visual encoder with activating each visage-style feature. Fig. 3 indicates that each visage-style feature, s^1, s^2, or s^3, sees different part of the face, meaning that each style feature is affected by different facial attributes.

3 LRW subject id generation

we clustered and labeled speaker information of a large-scale unconstrained audio-visual dataset, LRW [2]. Then, we split the train and test set without speaker overlapping to distinguish from the original splits of the dataset. Specifically, the speaker id of the LRW is labeled with similar pipeline of [1]: feature extraction, clustering, face verification and identification, and manual correction.

Firstly, we employ a powerful face recognition system, ArcFace [3], to represent the speaker feature of a video. With ResNet-101 [5] model pre-trained on MS-Celeb-1M [4], 5 frames of each video from the LRW are randomly selected for feature extraction. Then, the video-level speaker representation is obtained by averaging that of 5 frames embedded through the pre-trained face recognition model.

With the obtained video-level speaker representations, we cluster speakers through face identification between videos and clusters. If the cosine similarity between a given video and all clusters is lower than a threshold, a new cluster is created for the video. Otherwise, the video is assigned to a cluster that corresponds to the highest similarity. The speaker feature representing the cluster is updated with a new assigned video.

Next, face verification and identification are performed. Due to the imperfection of clustering algorithms, having false positive samples are inevitable. To minimize the error, we should remove the false positive samples that different speakers are assigned to one cluster. To this end, face verification is performed between all samples in a cluster. Then, face identification is proceeded between clusters to deal with the multiple clusters of one speaker that should be merged. To represent the cluster-level speaker feature, the video-level speaker representations of all videos in the cluster are averaged. Each cluster is compared with the other clusters, and it is merged with multiple top similarity clusters above a threshold. Finally, manual correction is performed for existing multiple clusters that should be merged.

Total $17,580$ subjects are labeled, $17,560$ for training, 20 for validation, and 20 for testing. Each split contain same about of class, 500. The number of videos are $480,378$, $29,918$, and $29,923$ for training, validation, and testing, respectively. We provide the train, validation, and test splits in ***LRW_train.txt***, ***LRW_val.txt***, and ***LRW_test.txt***, respectively. The txt files contain subject id with the location of the following video. The detailed explanation of the dataset is described in https://github.com/ms-dot-k/LRW_ID.

4 Architectural details

In this section, we describe the detailed architecture of each module in the proposed method. The architectures of visual encoder Φ_{VE}, visage-style encoder Φ_{VS}, VS-synthesizer Ψ_{VS}, visual-id classifier φ_v, audio-id classifier φ_a, discriminator, and postnet are illustrated in Table 5, 6, 7, 8, 9, 10, and 11, 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 mel-spectrogram is represented with 80 mel-spectral dimension. When converting audio to the mel-spectrogram, we use window size of 800 and hop size of 160 for 25fps videos, and window size of 532 and hop size of 133 for 30fps videos, in order to make the length of the mel-spectrogram 4 times longer than that of the video frames.

Table 5. Architecture of visual encoder

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 × $\frac{H}{2}$ × $\frac{W}{2}$ × 64
Max Pool 3D	1 × 3 × 3 / - / [1, 2, 2]	T × $\frac{H}{4}$ × $\frac{W}{4}$ × 64
ResBlock 2D	3 × 3 / 64 / [1, 1]	T × $\frac{H}{4}$ × $\frac{W}{4}$ × 64
ResBlock 2D	3 × 3 / 64 / [1, 1]	T × $\frac{H}{4}$ × $\frac{W}{4}$ × 64
ResBlock 2D	3 × 3 / 128 / [2, 2]	T × $\frac{H}{8}$ × $\frac{W}{8}$ × 128
ResBlock 2D	3 × 3 / 128 / [1, 1]	T × $\frac{H}{8}$ × $\frac{W}{8}$ × 128
ResBlock 2D	3 × 3 / 256 / [2, 2]	T × $\frac{H}{16}$ × $\frac{W}{16}$ × 256
ResBlock 2D	3 × 3 / 256 / [1, 1]	T × $\frac{H}{16}$ × $\frac{W}{16}$ × 256
ResBlock 2D	3 × 3 / 512 / [2, 2]	T × $\frac{H}{32}$ × $\frac{W}{32}$ × 512
ResBlock 2D	3 × 3 / 512 / [1, 1]	T × $\frac{H}{32}$ × $\frac{W}{32}$ × 512
Avg Pool 2D	-	T × 512

Table 6. Architecture of visage-style encoder

Visage-Style Encoder: input size T × C			
Style	**Layer**	**Hidden dim**	**Output dimensions**
s^1, s^2, s^3	Bi-GRU	512	T × 1024
s^1, s^2, s^3	Linear	1024 × 512	T × 512
s^1	Avg Pool	-	1 × 512
s^2, s^3	Linear	512 × 256	T × 256
s^2	Avg Pool	-	1 × 256
s^3	Linear	256 × 128	T × 128
s^3	Avg Pool	-	1 × 128

Table 7. Architecture of VS-synthesizer

Generator: input size 20 × T × (D+128)			
Layer	**Filter size / number / stride**	**Output dimensions**	**Norm**
ResBlock 2D	5 × 5 / 256 / [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	3 × 3 / 128 / [1, 1]	20 × T × 128	AdaIN (s^1)
ResBlock 2D	3 × 3 / 128 / [1, 1]	20 × T × 128	AdaIN (s^1)
ResBlock 2D	3 × 3 / 128 / [1, 1]	20 × T × 128	AdaIN (s^1)
ResBlock 2D	3 × 3 / 64 / [2, 2]	40 × 2T × 64	AdaIN (s^2)
ResBlock 2D	3 × 3 / 64 / [1, 1]	40 × 2T × 64	AdaIN (s^2)
ResBlock 2D	3 × 3 / 64 / [1, 1]	40 × 2T × 64	AdaIN (s^2)
ResBlock 2D	3 × 3 / 32 / [2, 2]	80 × 4T × 32	AdaIN (s^3)
ResBlock 2D	3 × 3 / 32 / [1, 1]	80 × 4T × 32	AdaIN (s^3)
ResBlock 2D	3 × 3 / 32 / [1, 1]	80 × 4T × 32	AdaIN (s^3)
Conv 2D	1 × 1 / 1 / [1, 1]	80 × 4T × 1	

Table 8. Architecture of visual-id classifier

Visual identity classifier: input size 1 × 128		
Layer	**Hidden dim**	**Output dimensions**
Linear	128 × 128	1 × 128
Linear	128 × 64	1 × 64
Linear	64 × 64	1 × 64
Linear	64 × Num_class	1 × Num_class

Table 9. Architecture of audio-id classifier

Audio identity 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
Reshape	-	T × 20 * 256
Linear	256 * 20 × 512	T × 512
Conv 1D	3 / 512 / [1, 1]	T × 512
Conv 1D	4 / 256 / [2, 2]	$\frac{T}{2}$ × 256
Conv 1D	4 / 256 / [2, 2]	$\frac{T}{4}$ × 256
Avg Pool	-	1 × 256
Linear	256 × 128	1 × 128
Linear	128 × 64	1 × 64
Linear	64 × Num_class	1 × Num_class

Table 10. Architecture of discriminator

Discriminator: input size 80 × 4T × 1			
Condition	**Layer**	**Filter size / number / stride**	**Output dimensions**
C&UC	ResBlock 2D	5 × 5 / 32 / [2, 2]	40 × $\frac{T}{2}$ × 32
	ResBlock 2D	5 × 5 / 64 / [2, 2]	20 × $\frac{T}{4}$ × 64
	ResBlock 2D	5 × 5 / 128 / [2, 2]	10 × $\frac{T}{8}$ × 128
	ResBlock 2D	5 × 5 / 256 / [2, 2]	5 × $\frac{T}{16}$ × 256
	Conv 2D	5 × 5 / 256 / [1, 1]	1 × $\frac{T}{16}$-4 × 256
C	Avg Pool 2D	-	256
	Linear	1	1
UC	Cat w/ s^3	-	5 × $\frac{T}{16}$ × (256 + 128)
	Conv 2D	5 × 5 / 256 / [1, 1]	5 × $\frac{T}{16}$ × 256
	Conv 2D	5 × 5 / 256 / [1, 1]	1 × $\frac{T}{16}$-4 × 256
	Avg Pool 2D	-	256
	Linear	1	1

Table 11. 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

References

1. Anvari, Z., Athitsos, V.: A pipeline for automated face dataset creation from unlabeled images. In: Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments. pp. 227–235 (2019)
2. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Asian conference on computer vision. pp. 87–103. Springer (2016)
3. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
4. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European conference on computer vision. pp. 87–102. Springer (2016)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Hong, J., Kim, M., Park, S.J., Ro, Y.M.: Speech reconstruction with reminiscent sound via visual voice memory. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3654–3667 (2021)
7. Kim, M., Hong, J., Park, S.J., Ro, Y.M.: Multi-modality associative bridging through memory: Speech sound recollected from face video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 296–306 (2021)
8. Michelsanti, D., Slizovskaia, O., Haro, G., Gómez, E., Tan, Z.H., Jensen, J.: Vocoder-based speech synthesis from silent videos. In: Interspeech 2020. pp. 3530–3534 (2020)
9. Mira, R., Vougioukas, K., Ma, P., Petridis, S., Schuller, B.W., Pantic, M.: End-to-end video-to-speech synthesis using generative adversarial networks. arXiv preprint arXiv:2104.13332 (2021)
10. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: Learning individual speaking styles for accurate lip to speech synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13796–13805 (2020)
11. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 484–492 (2020)
12. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
13. Vougioukas, K., Ma, P., Petridis, S., Pantic, M.: Video-driven speech reconstruction using generative adversarial networks. arXiv preprint arXiv:1906.06301 (2019)