Seeing What You Said: Talking Face Generation Guided by a Lip Reading Expert – Supplementary Materials –

1. Audio Encoder Architecture

The details of the audio encoder to extract a local audio embedding given a spectrogram of a 0.2-second audio segment are shown in Tab. 1. Specifically, layers between two lines compose a block whose input and output are summed up via a residual connection. Meanwhile, batch normalization is applied after each of the convolutional layers.

Table 1. Audio encoder architecture. All parameters listed in the 'Filters' column are kernel sizes, output channels, strides, padding, and repetition of layers.

Layer Type	Filters	Output dim.
Conv 2D	$\{[3, 3], 32, [1, 1], 1\} \times 1$	32×80×16
Conv 2D	$\{[3, 3], 32, [1, 1], 1\} \times 2$	32×80×16
Conv 2D	$\{[3, 3], 64, [3, 1], 1\} \times 1$	64×27×16
Conv 2D	$\{[3, 3], 64, [1, 1], 1\} \times 2$	64×27×16
Conv 2D	$\{[3, 3], 128, [3, 3], 1\} \times 1$	128×9×6
Conv 2D	$\{[3, 3], 128, [1, 1], 1\} \times 2$	128×9×6
Conv 2D	$\{[3, 3], 256, [3, 2], 1\} \times 1$	256×3×3
Conv 2D	$\{[3, 3], 256, [1, 1], 1\} \times 1$	256×3×3
Conv 2D	$\{[3, 3], 512, [1, 1], 0\} \times 1$	512×1×1
Conv 2D	$\{[1, 1], 512, [1, 1], 0\} \times 1$	512×1×1

2. Audio Transformer Encoder Architecture

The audio transformer encoder is used to extract phoneme-level information in speech considering the global temporal dependency, namely global audio embeddings. We use speeches with varying lengths as the input to the transformer encoder. In practice, a speech is preprocessed to a spectra $S \in \mathbb{R}^{T \times F}$, where T and F are the numbers of frames and filter banks. Then, every 4 frames are stacked into one frame. Herein, F is fixed to 26.

There are 12 cascaded transformer blocks in the transformer encoder. The hidden layer dimension, feed-forward layer dimension and the number of attention heads are set to 768, 3072 and 12, respectively. Thus, the output of the transform encoder is denoted as $Z \in \mathbb{R}^{(T/4) \times 768}$. Afterwards, we take one frame of Z, which is timely aligned with the pose reference, as the global audio embedding.

3. Video Encoder Architecture

We use a video encoder to extract the identity and pose information to a united visual embedding from a concatenation (6×96×96) of an identity and a pose image. Tab. 2 illustrates the detailed architecture.

Table 2. Video encoder architecture. All parameters listed in the 'Filters' column are kernel sizes, output channels, strides, padding, and repetition of layers.

Layer Type	Filters	Output dim.
Conv 2D	$\{[7, 7], 16, [1, 1], 3\} \times 1$	16×96×96
Conv 2D	$\{[3, 3], 32, [2, 2], 1\} \times 1$	32×48×48
Conv 2D	$\{[3, 3], 32, [1, 1], 1\} \times 2$	32×48×48
Conv 2D	$\{[3, 3], 64, [2, 2], 1\} \times 1$	64×24×24
Conv 2D	$\{[3, 3], 64, [1, 1], 1\} \times 3$	64×24×24
Conv 2D	$\{[3, 3], 128, [2, 2], 1\} \times 1$	128×12×12
Conv 2D	$\{[3, 3], 128, [1, 1], 1\} \times 2$	128×12×12
Conv 2D	$\{[3, 3], 256, [2, 2], 1\} \times 1$	256×6×6
Conv 2D	$\{[3, 3], 256, [1, 1], 1\} \times 2$	256×6×6
Conv 2D	$\{[3, 3], 512, [2, 2], 1\} \times 1$	512×3×3
Conv 2D	$\{[3, 3], 512, [1, 1], 1\} \times 1$	512×3×3
Conv 2D	$\{[3, 3], 512, [1, 1], 0\} \times 1$	512×1×1
Conv 2D	$\{[1, 1], 512, [1, 1], 0\} \times 1$	512×1×1

4. Generator Architecture

The details of the generator to synthesize a face image based on the concatenated audio and video embedding are provided in Tab. 3:

Besides, the skip connection like Unet [1, 2] is applied. Particularly, hidden features in the generator are concatenated with hidden features in the video encoder with the

Table 3. Generator architecture. All parameters listed in the 'Filters' column for Conv2D are kernel sizes, output channels, strides, padding, and repetition of layers. Conv 2D T. means 2D transposed convolutional layers which has an extra parameter called output padding, placed after the padding parameter.

Layer Type	Filters	Output dim.
Conv 2D	$\{[1, 1], 512, [1, 1], 0\} \times 1$	512×1×1
Conv 2D T. Conv 2D	$\{[3, 3], 512, [2, 2], 0, 0\} \times 1$ $\{[3, 3], 512, [1, 1], 1\} \times 1$	512×3×3 512×3×3
Conv 2D T. Conv 2D	$\{[3, 3], 512, [2, 2], 1, 1\} \times 1$ $\{[3, 3], 512, [1, 1], 1\} \times 2$	512×6×6 512×6×6
Conv 2D T. Conv 2D	$\{[3, 3], 384, [2, 2], 1, 1\} \times 1$ $\{[3, 3], 384, [1, 1], 1\} \times 2$	384×12×12 384×12×12
Conv 2D T. Conv 2D	$\{[3, 3], 256, [2, 2], 1, 1\} \times 1$ $\{[3, 3], 256, [1, 1], 1\} \times 2$	256×24×24 256×24×24
Conv 2D T. Conv 2D	$\{[3, 3], 128, [2, 2], 1, 1\} \times 1$ $\{[3, 3], 128, [1, 1], 1\} \times 2$	128×48×48 128×48×48
Conv 2D T. Conv 2D	$\{[3, 3], 64, [2, 2], 1, 1\} \times 1$ $\{[1, 1], 64, [1, 1], 1\} \times 2$	64×96×96 64×96×96
Conv 2D Conv 2D	$\{[3, 3], 32, [1, 1], 1\} \times 1$ $\{[1, 1], 3, [1, 1], 0\} \times 1$	32×96×96 3×96×96

same shape.

5. Discriminator Architecture

The details of the discriminator to penalize unrealistic synthesized face images are provided in Tab. 4. The discriminator only takes the lower half of faces as inputs.

6. Qualitative Ablation Study on Contrastive Learning

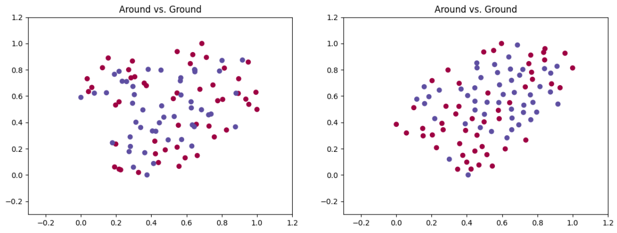
Our experiments have confirmed that contrastive learning is effective in lip-speech synchronization, which also improves reading intelligibility. In this section, we visualize audio embeddings of the pairs of 'Around' and 'Ground' which is one of the most frequently confused word pairs [3].

As shown in Fig. 1, although audio embeddings of two words by the TalkLip ($l + c$) are still not separated, the lower half of Fig. 1b is mainly composed by red points, which is much better than the TalkLip (l).

Besides, we provide some figures to conduct a qualitative analysis. As shown in Fig. 2, it is observed that the TalkLip ($l + c$) is better than Prop. (l) with the help of contrastive learning. Especially the fourth image (in the blue box) of the TalkLip (l) is a little ahead of the ground truth. The fourth image of the TalkLip ($l + c$) is more synchronized than that of the TalkLip (l) with the ground truth.

Table 4. Audio encoder architecture. All parameters listed in the 'Filters' column are kernel sizes, output channels, strides, padding, and repetition of layers.

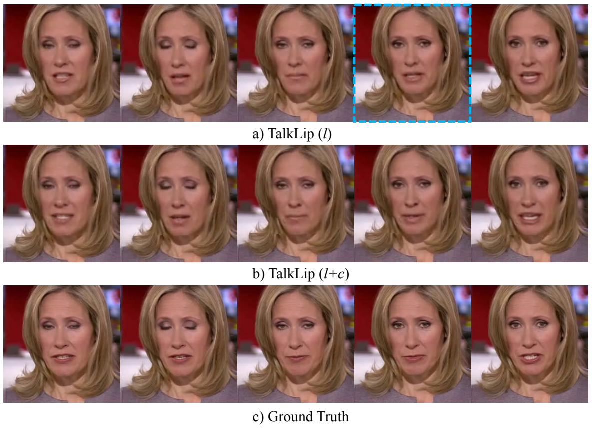
Layer Type	Filters	Output dim.
Conv 2D	$\{[7, 7], 32, [1, 1], 3\} \times 1$	32×48×96
Conv 2D Conv 2D	$\{[5, 5], 64, [1, 2], 1\} \times 1$ $\{[5, 5], 64, [1, 1], 2\} \times 1$	64×48×48 64×48×48
Conv 2D Conv 2D	$\{[5, 5], 128, [2, 2], 2\} \times 1$ $\{[5, 5], 128, [1, 1], 2\} \times 1$	128×24×24 128×24×24
Conv 2D Conv 2D	$\{[5, 5], 256, [2, 2], 2\} \times 1$ $\{[5, 5], 256, [1, 1], 2\} \times 1$	256×12×12 256×12×12
Conv 2D Conv 2D	$\{[5, 5], 512, [2, 2], 2\} \times 1$ $\{[5, 5], 512, [1, 1], 2\} \times 1$	512×6×6 512×6×6
Conv 2D Conv 2D	$\{[3, 3], 512, [2, 2], 1\} \times 1$ $\{[3, 3], 512, [1, 1], 1\} \times 1$	512×3×3 512×3×3
Conv 2D Conv 2D	$\{[3, 3], 512, [1, 1], 0\} \times 1$ $\{[1, 1], 512, [1, 1], 0\} \times 1$	512×1×1 512×1×1



(a) TalkLip (l)

(b) TalkLip (l+c)

Figure 1. The t-SNE visualization of audio embeddings correspond to 'Around' (red) and 'Ground' (blue).



a) TalkLip (l)

b) TalkLip (l+c)

c) Ground Truth

Figure 2. Snapshots of the generated talking face videos to demonstrate the benefit of the contrastive learning.

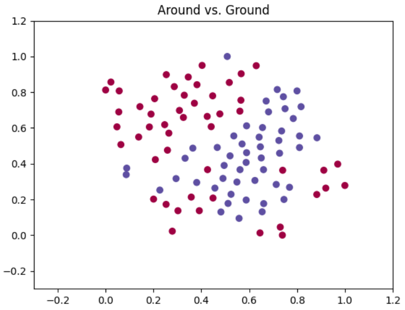
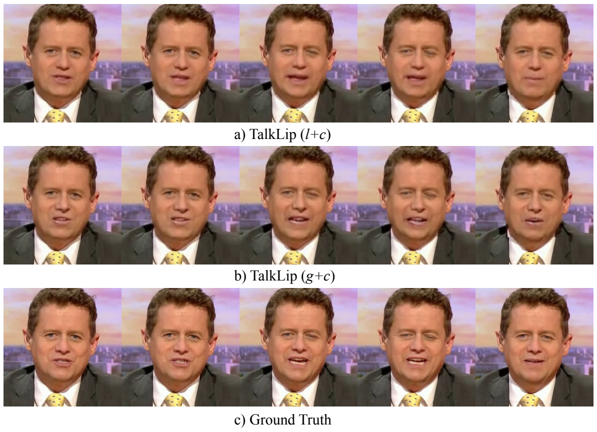


Figure 3. The t-SNE visualization of global audio embeddings correspond to 'Around' (red) and 'Ground' (blue).



a) TalkLip (l+c)

b) TalkLip (g+c)

c) Ground Truth

Figure 4. Snapshots of the generated talking face videos to demonstrate the benefit of using the transformer encoder which extracts the global audio embedding.

7. Qualitative Ablation Study on Global Audio Embedding

To show a better representation of global audio embeddings in phoneme-level information, we visualize their distributions of 'Around' and 'Ground' in Fig. 3. It is observed that global audio embeddings are more separable than local audio embeddings as shown in Fig. 1. Besides, We show an image comparison between the TalkLip ($g + c$) and the TalkLip ($l + c$) as Fig. 4. It is observed that the lip movement of the TalkLip ($g+c$) is fuller than the TalkLip ($l+c$), which confirms the benefit of the global audio embedding.

8. LSE-D

LSE-D [2] is another metric to measure lip-speech synchronization. We provide a comparison of LSE-D in Tab. 5. We can observe that LSE-D also confirms the SOTA performance of TalkLip (g + c) on lip-speech synchronization.

Table 5. LSE-D results on LRW and LRS2. Performances of methods with * are collected from [4] which are trained using the whole LRS2 dataset (224 hours) while our methods are trained by the LRS2 dataset (29 hours).

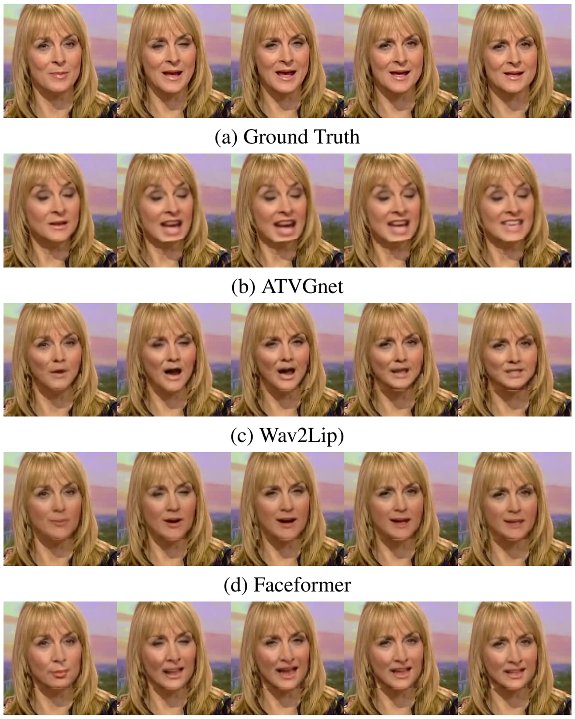
	LRW	LRS2
Ground Truth	6.97	6.45
ATVGnet	8.56	8.65
Wav2Lip	7.01	6.58
Faceformer	8.00	7.80
PC-AVS*	7.34	7.30
SyncTalkFace*	6.97	6.26
TalkLip (l)	7.00	6.63
TalkLip (l + c)	7.00	6.56
TalkLip (g)	6.75	6.01
TalkLip (g + c)	**6.51**	**6.00**

9. Limitation

In the Fig. 1 of the main body, it is observed that our methods do not help improve visual quality. The lip-reading loss does not direct a better visual quality since PSNR and SSIM of the TalkLip (l) and the **Base** w.o. $\mathcal{L}_{\text{lip}}$ are very close. Contrastive learning and the global audio encoder also do not boost visual quality as all four TalkLip nets show similar PSNR and SSIM. We will explore methods of improving visual quality in our further work.

10. Qualitative Result

In this section, we provide more qualitative comparisons with 3 State-of-Arts methods: ATVGnet [5], Wav2Lip [2], Faceformer [6] to show the superiority of our proposal. Please see details in Fig. 5-8.



(a) Ground Truth

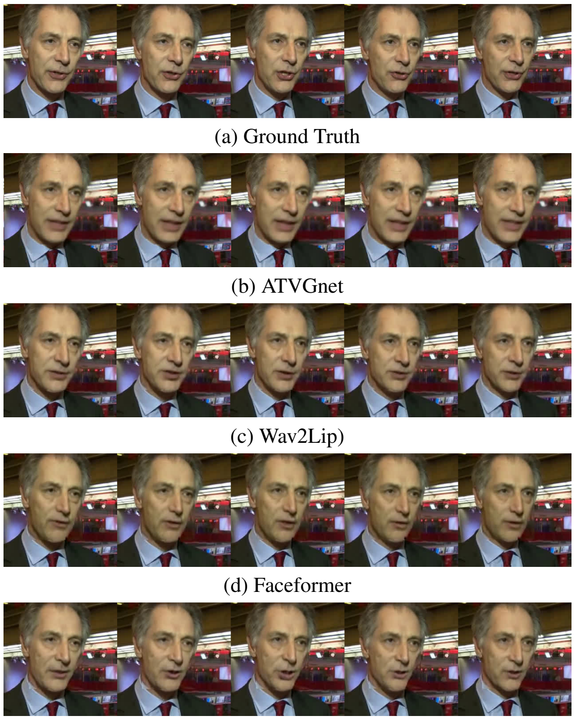
(b) ATVGnet

(c) Wav2Lip)

(d) Faceformer

(e) TalkLip ($g+c$)

Figure 5. Qualitative comparison.



(a) Ground Truth

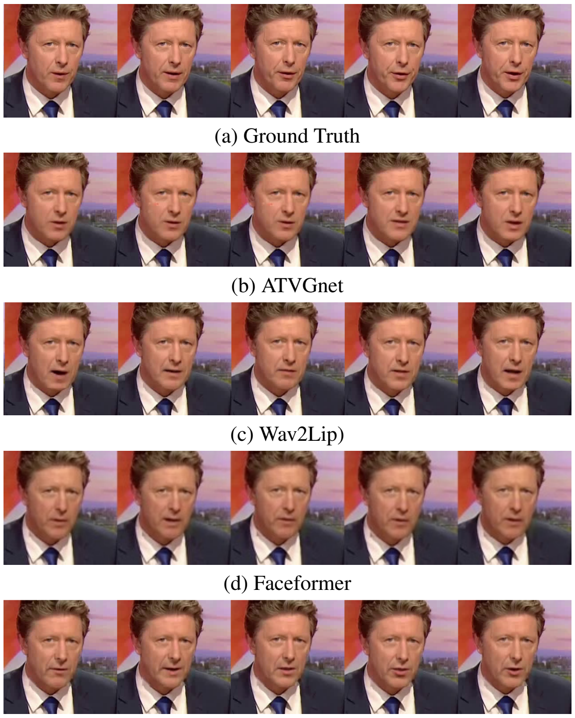
(b) ATVGnet

(c) Wav2Lip)

(d) Faceformer

(e) TalkLip ($g+c$)

Figure 6. Qualitative comparison.



(a) Ground Truth

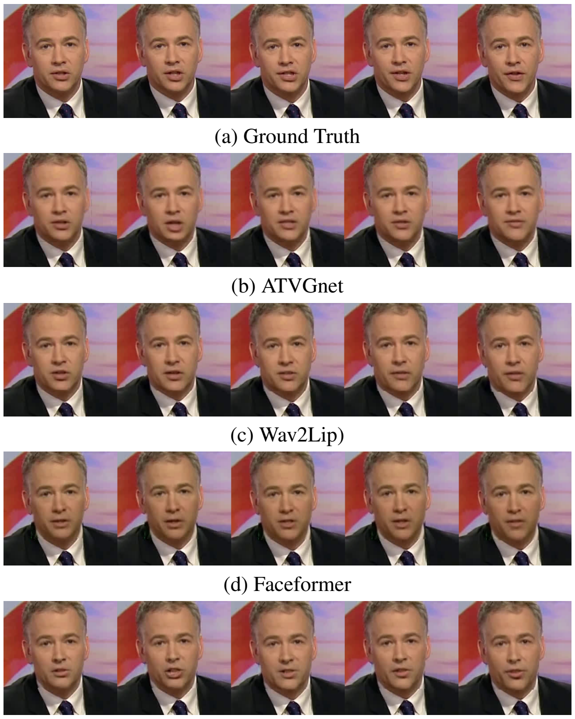
(b) ATVGnet

(c) Wav2Lip)

(d) Faceformer

(e) TalkLip ($g+c$)

Figure 7. Qualitative comparison.



(a) Ground Truth

(b) ATVGnet

(c) Wav2Lip)

(d) Faceformer

(e) TalkLip ($g+c$)

Figure 8. Qualitative comparison.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241. 1

[2] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "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*, 2020, pp. 484–492. 1, 3

[3] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 87–103. 2

[4] S. J. Park, M. Kim, J. Hong, J. Choi, and Y. M. R, "ynctalkface: Talking face generation with precise lip-syncing via audio-lip memory," in *Association for the Advancement of Artificial Intelligence*, 2022, pp. 234–778. 3

[5] L. Chen, H. Zheng, R. K. Maddox, Z. Duan, and C. Xu, "Sound to visual: Hierarchical cross-modal talking face video generation," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops*, 2019. 3

[6] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "Faceformer: Speech-driven 3d facial animation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 770–18 780. 3