

Mimic: Speaking Style Disentanglement for Speech-Driven 3D Facial Animation

Supplementary Material

We strongly recommend watching the supplementary video. We provide the full video and separately organized videos, respectively.

Abstract

This supplemental document contains five aspects. Section 6 shows the detailed architecture of our Mimic. Section 7 presents more details of datasets and implementations. Section 8 provides more experimental results and analysis. Section 9 describes more details of our user study. Section 10 describes our supplemental videos. We also provide the code to help readers better understand our implementation. The source code and trained model will also be released upon publication.

6 Network Architecture Details

To improve the reproducibility of our method, we further illustrate the detailed architectures of our Mimic in Table 5. Note that the parameters of the audio feature extractor (TCN) and transformer encoder in our audio encoder are initialized with the pre-trained wav2vec 2.0 weights¹. The parameters of the audio feature extractor are fixed during training.

7 More Details of Datasets and Implementations

More Details of Datasets

We use 3D-HDTF, VOCASET (Cudeiro et al. 2019), and BIWI (Fanelli et al. 2010) to evaluate our method qualitatively and quantitatively. The detailed descriptions of the three datasets are as follows.

3D-HDTF We construct a large-scale dataset called 3D-HDTF for speech-driven 3D facial animation, which is based on a large in-the-wild high-resolution audio-visual dataset HDTF (Zhang et al. 2021), consisting of about 362 different videos for 15.8 hours, and the resolution of the origin video is 720P or 1080P. To obtain 3D supervision, we generate pseudo ground truth 3D mesh data by integrating an off-the-shelf monocular 3D face reconstruction method named SPECTRE (Filntisis et al. 2022), which incorporates a perceptual lip reading loss to guide the better reconstruction of mouth movements.

¹<https://huggingface.co/facebook/wav2vec2-base-960h>

SPECTRE predicts the shape, expression, and pose parameters of FLAME (Li et al. 2017), which is a parametric 3D head model. FLAME uses linear transformations to describe identity and expression-dependent shape variations, as well as standard linear blend skinning (LBS) to model neck, jaw, and eyeball rotations. It is defined as a function $M(\beta, \theta, \psi) \rightarrow (\mathbf{V}, \mathbf{F})$ that given parameters for identity shape $\beta \in \mathbb{R}^{|\beta|}$, expression $\psi \in \mathbb{R}^{|\psi|}$, and pose $\theta \in \mathbb{R}^{|\theta|}$, outputs a 3D mesh with vertices $\mathbf{V} \in \mathbb{R}^{n_c \times 3}$ and triangles $\mathbf{F} \in \mathbb{R}^{n_f \times 3}$, where $n_c = 5023$ and $n_f = 9976$.

We employ SPECTRE trained on HDTF to obtain more accurate FLAME parameters and get meshes by FLAME forward pass. To align all frames in the “zero pose”, we set the pose parameters describing neck rotations as zero. To get a template mesh of a subject, we feed the shape parameters of the first frame of the subject, the zero-valued expression parameters, and the zero-valued pose parameters into the FLAME model. After filtering out some unavailable samples, such as face detection failures, audio-visual misalignment, etc., we obtain 220 mesh sequences of 160 identities with over 3k unique sentences and corresponding 160 template meshes. All sequences are processed at 25fps.

Compared with currently available 3D facial animation datasets, such as VOCASET and BIWI, our 3D-HDTF contains highly diversified subjects and corpus, holding extensive speaking and rich speech content, which builds a more informative facial motion space and better evaluates the effectiveness of our disentanglement method. Such a large-scale dataset enables the training of high-fidelity, expressive, and generalizable face animation models. Therefore, we conduct extensive qualitative and quantitative experiments on 3D-HDTF.

BIWI BIWI (Fanelli et al. 2010) is a 3D audio-visual corpus of affective speech and corresponding dense dynamic 3D face geometries. It comprises two parts, one with emotions and the other devoid of them, with 40 unique sentences uttered by 14 subjects. In total, it consists of 1109 sequences captured at 25 fps, 4.67 seconds long on average. All 3D face geometries are registered with 23370 vertices. We follow the data splits in recent works (Fan et al. 2022; Xing et al. 2023) and use the emotional subset. Specifically, the training set (BIWI-Train) contains 192 sentences, while the validation set (BIWI-Val) contains 24 sentences. There are two testing

Module	Input→Output	Operation
Style Encoder	$\mathbf{M}(T, V, 3) \rightarrow \mathbf{M}(T, V \times 3)$	Reshape
	$\mathbf{M}(T, V \times 3) \rightarrow s(128)$	$L(128) \rightarrow [C(3, 1, 1, 128) \rightarrow LN \rightarrow ReLU] \times 5 \rightarrow LN \rightarrow L(128) \rightarrow Drop \rightarrow PE \rightarrow T_{enc}(128, 256, 4, 4) \rightarrow L(128) \rightarrow MeanPool$
Content Encoder	$\mathbf{M}(T, V, 3) \rightarrow \mathbf{M}(T, V \times 3)$	Reshape
	$\mathbf{M}(T, V \times 3) \rightarrow s(128)$	$L(128) \rightarrow [C(3, 1, 1, 128) \rightarrow LN \rightarrow ReLU] \times 5 \rightarrow LN \rightarrow L(128) \rightarrow Drop \rightarrow PE \rightarrow T_{enc}(128, 256, 4, 4) \rightarrow L(128)$
Audio Encoder	$\mathcal{X}(L) \rightarrow a'(T_a, 512)$	$C(10, 5, 0, 512) \rightarrow GN \rightarrow GeLU \rightarrow [C(3, 2, 0, 512) \rightarrow GN \rightarrow GeLU] \times 7$
	$a'(T_a, 512) \rightarrow a(T, 128)$	$C(15, 2, 7, 512) \rightarrow LN \rightarrow L(768) \rightarrow Drop \rightarrow PE \rightarrow T_{enc}(768, 3072, 12, 12) \rightarrow L(128)$
Motion Decoder	$\tilde{\mathbf{M}}_{past}(T, V \times 3) \rightarrow \mathbf{m}_{past}(T, 128)$	$L(128)$
	$\mathbf{m}_{past}(T, 128) \rightarrow \mathbf{m}(T, 128)$	$PPE \rightarrow [MSA(128, 4) \rightarrow SALN(s) \rightarrow MCA(128, 4, a c) \rightarrow SALN(s) \rightarrow FF(128, 256) \rightarrow SALN(s)] \times 1$
	$\mathbf{m}(T, 128) \rightarrow \tilde{\mathbf{M}}(T, V \times 3)$	$L(V \times 3)$

Table 5: Illustration of detailed architectures. We set the d_a , d_c , and d_a in the main paper as 128. $C(k, s, p, n)$ denotes a 1D convolutional layer with kernel size k , stride size s , padding size p , and output channels of n . $T_{enc}(d_1, d_2, h, l)$ denotes a transformer encoder layer with basic channels of d_1 , forward channels of d_2 , self-attention head number of h , and layers of l . $L(n)$ denotes a linear layer with output channels of n . Drop denotes the dropout operation. PE and PPE denote the positional encoding (Vaswani et al. 2017) and periodic positional encoding (Fan et al. 2022) operation. MeanPool denotes a mean pooling along the temporal axis. $MSA(d, h)$ denotes a multi-head self-attention with basic channels of d and head number of h . $MCA(d, h, f)$ denotes a multi-head cross-attention with basic channels of d , head number of h and input features f . $SALN(s)$ denotes style-adaptive layer normalization with input style code s . $FF(d_1, d_2)$ denotes a feed forward layer with basic channels of d_1 and forward channels of d_2 . $V = 5023$ for 3D-HDTF and VOCASET, and $V = 23370$ for BIWI. We set $T = 150$ (6s) for 3D-HDTF and a T that needs to be determined by the length of the sampled input sequence for both VOCASET and BIWI.

sets, in which BIWI-Test-A includes 24 sentences spoken by six seen subjects, and BIWI-Test-B contains 32 sentences spoken by eight unseen subjects. In our study, we use the BIWI-Test-B for qualitative evaluation.

VOCASET VOCASET consists of 480 paired audio-visual sequences recorded from 12 subjects, which are captured at 60 fps, and each sequence is about 4 seconds long. It contains 255 unique sentences, some of which are shared across speakers. Each 3D face mesh is registered to the FLAME topology with 5023 vertices. We adopt the same training (VOCA-Train), validation (VOCA-Val), and testing (VOCA-Test) splits as recent works (Fan et al. 2022; Xing et al. 2023) for fair comparisons.

Implementations of Baseline Methods

As mentioned in the main paper, we compare our method with five state-of-the-art methods, VOCA (Cudeiro et al. 2019), MeshTalk (Richard et al. 2021), FaceFormer (Fan et al. 2022), CodeTalker (Xing et al. 2023), and Imitator (Thambiraja et al. 2022). For VOCA, we use the official code² to train and test on 3D-HDTF and BIWI, and test the released model on VOCASET. For MeshTalk, we train and test it using the official code³ on the three datasets. For FaceFormer⁴ and CodeTalker⁵, we use the official code to train and test on 3D-HDTF and test the released model on both VOCASET and BIWI. For Imitator, we implement it to the best of our understanding. We train Imitator on the three datasets and perform the two-stage adaptation with a 10-second style reference video of each target subject.

²<https://github.com/TimoBolkart/voca>

³<https://github.com/facebookresearch/meshtalk>

⁴<https://github.com/EvelynFan/FaceFormer>

⁵<https://github.com/Doubiuu/CodeTalker>

8 More Experimental Results and Analysis

More Results and Analysis of Constraints

As described in the main paper, our method achieves the best performance with all four constraints, demonstrating facilitation for disentangled representation learning by our constraints. We further investigate the impact of the auxiliary style classifier C_s , auxiliary inverse classifier C_c , and content contrastive loss \mathcal{L}_{con} on two latent spaces. We conduct ablation studies on 3D-HDTF-Test-B and visualize the latent spaces of ablation results in Figure 7. Comparing Figure 7 (b) with (a), we observe that incorporating an auxiliary style classifier brings the style codes of the same subject closer to a common clustering center while pulling away style codes from different subjects, contributing to the construction of identity-related style space. Moving to Figure 7 (c), we noticed that the content codes of the same subject tend to be distributed in neighboring regions, suggesting that they exhibit similar identity-related information. From (c) to (d) in Figure 7, the introduced auxiliary inverse classifier helps alleviate the clustering effect among content codes, leading to a more dispersed distribution. Furthermore, with the content contrastive loss, the distributions of content codes and audio features are pulled closer together, resulting in the content space containing more semantic content information, which further reduces the clustering of the content codes, as shown in Figure 7 (e). With both the auxiliary inverse classifier and content contrastive loss, the content space contains minimal identity-related information, thus minimizing overlap with the style space, ultimately enabling effective disentanglement of speaking style and semantic content.

Impact of SALN

We introduce the style-adaptive layer normalization (SALN) (Min et al. 2021) to incorporate the style code into our decoder. To verify the effectiveness of SALN, we compare it

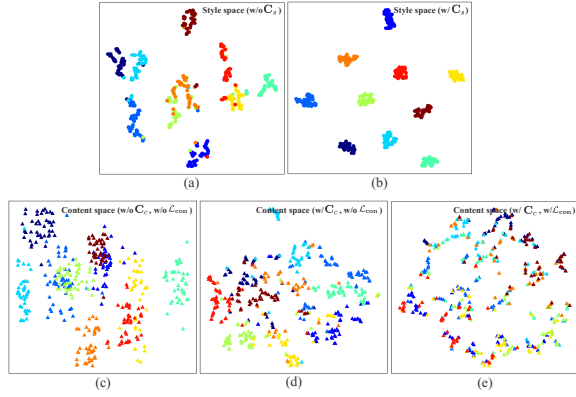


Figure 7: Impact of C_s on style space (a, b), and impact of C_c and L_{con} on content space (c, d, e). Different colors indicate the latent codes of different subjects.

Method	FVE ↓ ($\times 10^{-6}$ mm)	LVE ↓ ($\times 10^{-5}$ mm)	LDTW ↓ ($\times 10^{-4}$)	LDD ↓ ($\times 10^{-5}$ mm)	SCS ↑
Ours (SALN)	0.551	3.20	6.82	0.89	0.995
Ours (add)	0.557	3.49	6.88	0.91	0.992
Ours (cat)	0.559	3.54	6.94	0.92	0.990

Table 6: Impact of SALN on 3D-HDTF-Test-A.

with the operations commonly used in recent works, such as adding (Fan et al. 2022; Xing et al. 2023) and concatenating (Cudeiro et al. 2019; Thambiraja et al. 2022). We test our framework trained with these operations on 3D-HDTF-Test-A and show results in Table 6. It can be seen that our framework with SALN achieves the best performance among all metrics, demonstrating the superiority of SALN. It may be due to the fact that simple operations like adding or concatenating can hardly inject the rich style information within style code into the decoding process.

Impact of Style Reference Sequence Length

To explore the impact of style reference sequence length, we conduct an experiment on 3D-HDTF-Test-B. We sample sequences of different lengths (1-10s) as our style reference sequences, respectively, and calculate the style-related metrics (LDD and SCS) for the inference results. As shown in Figure 8, we observed that the performance increases as the sequence length becomes longer, up until the sequence length reaches approximately 6 seconds. This may be due to the fact that the sequence length we used for training is 6s. Despite a slight decrease in performance when reducing the sequence length, our model still outperforms Imitator (Thambiraja et al. 2022), which uses a 10-second sequence length for style adaptation.

Video-driven 3D Facial Animation

Benefiting from our style-content disentanglement, we can also achieve video-driven 3D facial animation, which requires a driving video clip and a short style reference sequence as inputs. The input driving video can provide se-

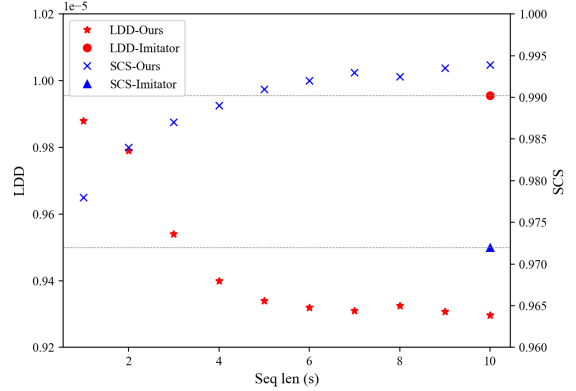


Figure 8: Impact of the style reference sequence length.

semantic content information just like the driving speech. We provide demos of 3D-HDTF in our supplemental videos.

9 More Details of User Study

As mentioned in the main paper, we collect 300 entries each on 3D-HDTF-Test-A, VOCA-Test, and BIWI-Test-B, and 450 entries on 3D-HDTF-Test-B. In total, we obtain 1350 entries. In this study, 30 participants with good vision and hearing ability complete the evaluation successfully. We ensure that the entries of each dataset were equally distributed to each participant. As a result, the user study interface shows 45 video pairs for each participant. For evaluating the perceptual lip sync and realism, the participant is instructed to judge the videos twice with the following two questions, respectively: “Comparing the lips of two faces, which one is more in sync with the audio?” and “Comparing the two full faces, which one looks more realistic?”, as shown in Figure 9. For evaluating the speaking style, the participant is instructed to judge the videos with the question: “Comparing the speaking style (including the amplitude of mouth opening and closing, the dimensionality of pouting, etc.) of the last two faces, which one is more consistent with the first video?”, as shown in Figure 10. To avoid any selection bias, the order of all methods for comparison is random for each pair. We filter out those comparison results completed in less than two minutes to remove the impact of random selection.

10 Video Comparison

To better evaluate the qualitative results produced by competitors (VOCA (Cudeiro et al. 2019), MeshTalk (Richard et al. 2021), FaceFormer (Fan et al. 2022), CodeTalker (Xing et al. 2023), and Imitator (Thambiraja et al. 2022)) and our Mimic, we provide a supplemental video for demonstration and comparison. The supplemental video contains the following results:

- Qualitative test results on 3D-HDTF-Test-B
- Qualitative test results on VOCA-Test
- Qualitative test results on BIWI-Test-B

- Comparison to SOTA methods on 3D-HDTF-Test-A
- Comparison to Imitator on 3D-HDTF-Test-B
- Comparison to SOTA methods on VOCA-Test
- Comparison to SOTA methods on BIWI-Test-B
- Comparison to previous methods on the speech from supplementary videos of previous methods
- Qualitative test results of speaking style interpolation
- Qualitative test results of video-driven facial animation
- Qualitative test results of different languages (German)
- Qualitative test results of different languages (Korean)

Separately organized videos are also provided for ease of watching.

References


- Cudeiro, D.; Bolkart, T.; Laidlaw, C.; Ranjan, A.; and Black, M. J. 2019. Capture, Learning, and Synthesis of 3D Speaking Styles. In *CVPR*, 10101–10111.
- Fan, Y.; Lin, Z.; Saito, J.; Wang, W.; and Komura, T. 2022. Faceformer: Speech-Driven 3D Facial Animation with Transformers. In *CVPR*, 18770–18780.
- Fanelli, G.; Gall, J.; Romsdorfer, H.; Weise, T.; and Van Gool, L. 2010. A 3D Audio-Visual Corpus of Affective Communication. *IEEE Transactions on Multimedia*, 12(6): 591–598.
- Filintisis, P. P.; Retsinas, G.; Paraperas-Papantoniou, F.; Katsamanis, A.; Roussos, A.; and Maragos, P. 2022. Visual Speech-Aware Perceptual 3D Facial Expression Reconstruction from Videos. *arXiv preprint arXiv:2207.11094*.
- Li, T.; Bolkart, T.; Black, M. J.; Li, H.; and Romero, J. 2017. Learning a Model of Facial Shape and Expression from 4D Scans. *ACM Transactions on Graphics (TOG)*, 36(6): 194–1.
- Min, D.; Lee, D. B.; Yang, E.; and Hwang, S. J. 2021. Meta-StyleSpeech: Multi-Speaker Adaptive Text-to-Speech Generation. In *ICML*, 7748–7759.
- Richard, A.; Zollhöfer, M.; Wen, Y.; De la Torre, F.; and Sheikh, Y. 2021. MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement. In *ICCV*, 1173–1182.
- Thambiraja, B.; Habibie, I.; Aliakbarian, S.; Cosker, D.; Theobalt, C.; and Thies, J. 2022. Imitator: Personalized Speech-driven 3D Facial Animation. *arXiv preprint arXiv:2301.00023*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All you Need. *Advances in Neural Information Processing Systems*, 30.
- Xing, J.; Xia, M.; Zhang, Y.; Cun, X.; Wang, J.; and Wong, T.-T. 2023. CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior. In *CVPR*, 12780–12790.
- Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021. Flow-guided One-shot Talking Face Generation with a High-resolution Audio-visual Dataset. In *CVPR*, 3661–3670.

Instructions

Please watch these short videos (duration 4~10s) of two or three animated talking heads and answer the following two or three questions.

Reminder: Please turn on the sound on your computer while you are watching the videos.


1 Please watch the video and answer the questions.



	The left	The right
Q1. Comparing the lips of the two faces, which one is more in sync with the audio?	<input type="radio"/>	<input type="radio"/>
Q2. Comparing the two full faces, which one looks more realistic?	<input type="radio"/>	<input type="radio"/>

Figure 9: User study interface for evaluating the perceptual lip sync and realism.

41 Please watch the video and answer the question.



	The second	The third
Q1. Comparing the speaking style (including the amplitude of mouth opening and closing, the dimensionality of pouting, etc.) of the last two faces, which one is more consistent with the first video?	<input type="radio"/>	<input type="radio"/>

Figure 10: User study interface for evaluating the speaking style.