Speech4Mesh: Speech-Assisted Monocular 3D Facial Reconstruction for Speech-Driven 3D Facial Animation
—— Supplementary Material ——

Shan He[1,2], Haonan He[2], Shuo Yang[2], Xiaoyan Wu[2], Pengcheng Xia[2],
Bing Yin[2], Cong Liu[2], Lirong Dai[1], Chang Xu[3]
[1]University of Science and Technology of China, [2]iFLYTEK Research, [3]University of Sydney
{shanhe2, hnhe, shuoyang7, xywu10, pcxia, bingyin, congliu2}@iflytek.com
lrdai@ustc.edu.cn, c.xu@sydney.edu.au

In this supplementary material, we provide detailed implementation information of the speech-assisted monocular 3D facial reconstruction module, additional experimental results, and a description of the supplementary video.

1. Implementation Details of Speech-Assisted Monocular 3D Face Reconstruction

We trained our monocular 3D face reconstruction network with $\lambda_{con} = 0.025$, $\lambda_{lmk} = 50$, $\lambda_{eye} = 25$, $\lambda_{lp} = 25$, $\lambda_{pho} = 25$ and $\lambda_{emo} = 0.5$. λ_{reg} consists of regularization terms of jaw $\lambda_{\theta jaw} = 200$ and expression $\lambda_{\psi} = 1e-3$. For the weighted landmark loss, we set different weights for key points in different areas. Weights of landmarks related to the jaw were set to 2, weights of mouth corners and nose tips landmarks were set to 3, other mouth and nose landmarks were weighted by a factor of 1.5, and the remaining landmarks had the same weight of 1.0.

For contrastive learning, we extracted information contained in the speech by a fixed pre-trained wav2vec 2.0, which could learn powerful representations from speech. We took hidden states from the last layer of the Transformer architecture as the speech features. These features have a frequency of 49 Hz with a stride of about 20 ms between each sample. We linearly interpolated them to approximately double the video frame rate. For example, speech features were resampled to 60 Hz for the MEAD dataset in which the videos are 30 fps.

2. Additional Comparison Results on Speech-Assisted Monocular 3D Face Reconstruction

Full face error: To provide more comprehensive quantitative results, we compare our algorithm with DECA and EMOCA on the full face error. Tab. 1 shows that our model achieves a full face error much better than DECA and comparable to EMOCA. However, our method improves significantly on lip vertex error, which is more relevant for audio2mesh, as shown in the paper.

Loss	DECA	EMOCA	Ours
Face Vertex Error ↓	$2.686e^{-2}$	$\mathbf{2.401}e^{-2}$	$2.439e^{-2}$

Table 1: Full face error.

Comparison with SPECTRE: Since SPECTRE [3] is a contemporaneous work and has not been published, we did not compare it in the body text. Here, we compare the lip vertex error on the VOCASET dataset as in Tab. 2 and demonstrate some reconstruction results in Fig. 1. As illustrated in the results, our method exhibits superior performance over SPECTRE in capturing the dynamic motion of the mouth. Notably, the mouth shape as depicted in SPECTRE may appear over-exaggerated in certain instances, as shown in the bottom row of the figures.

Metric	Ours	SPECTRE
Lip Vertex Error ↓	$\mathbf{0.995}e^{-2}$	$1.261e^{-2}$

Table 2: Reconstruction error on VOCASET.

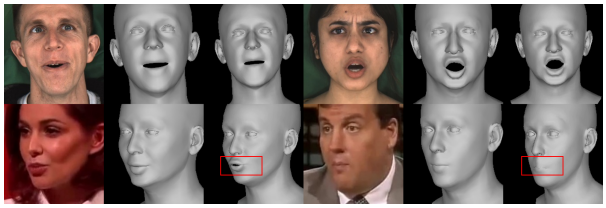

Figure 1: Comparison with SPECTRE on reconstruction results, left: RGB image; middle: ours, right: SPECTRE.

Comparison of Visualization Results with EMOCA: We would like to further compare our method with EMOCA [2], which can produce expressive results on emotional images (especially on the upper face region). However, in practice, it is not capable to capture the mouth geometry accurately (see Fig. 2) and might induce exaggerated expressions even for the images with neutral emotion. This attribute may render EMOCA unsuitable for the pre-training audio2mesh model.

Figure 2: Comparison with EMOCA on reconstruction results, left: RGB image; middle: EMOCA, right: ours.

3. Additional Ablation studies

Analysis of the impact of MEAD dataset on Monocular 3D Face Reconstruction: Because the MEAD dataset contains strong emotions and exaggerated expressions and because other methods, except for EMOCA, were not trained on the MEAD dataset, we further conducted an ablation study for the sake of fairness in comparison and to investigate if such strong emotions would affect the reconstruction result. We removed MEAD dataset from our training dataset and only trained our reconstruction model using VoxCeleb2 dataset, and the result is shown in Fig. 3. We can find that removing the MEAD dataset does not show significant degeneration in the reconstruction.

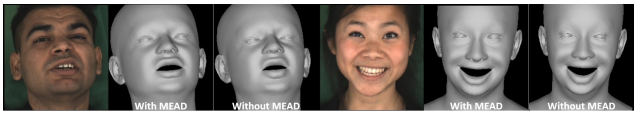

Figure 3: Reconstruction results w/wo training on MEAD.

Analysis of the impact of pre-training and fine-tuning on audio2mesh: For the training of our audio2mesh module, we first pre-trained it using 3D reconstructions obtained from our 3D face reconstruction module and then fine-tuned it using 4D scans from VOCASET. To investigate the impact of both the pre-training stage and the fine-tuning stage, we conduct an additional ablation study as shown in Tab. 3. Random is the result without either pre-training or fine-tuning. It indicates that both pre-training and fine-tuning play significant roles in our method.

Visualization Analysis of Contractive Loss's Impact on Monocular 3D Face Reconstruction: To further investigate the impact of our contrastive loss on the reconstruction results, we conducted an additional visualization analysis as shown in Fig. 4. Our findings indicate that training with contrastive loss can significantly improve the quality of 3D reconstruction, particularly in the mouth region. The third column of images was trained without contrastive loss, which performs inconspicuous lip deformations compared to the second column which was trained with contrastive loss.

Metric	Random	Pre-train	Fine-tune
Lip Vertex Error ↓	$1.849e^{-2}$	$8.491e^{-3}$	$\mathbf{4.863}e^{-3}$

Table 3: Ablation study on pre-train and fine-tune.

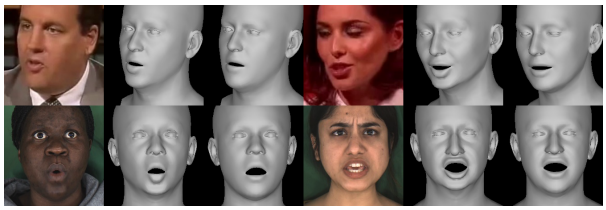

Figure 4: Ablation study on the contrastive loss w.r.t. the reconstruction quality.

We further visualized the t-SNE plots of the audio features and expression features extracted by the audio encoder E_s and the expression encoder E_e in our reconstruction module. We compared three different results: training with contrastive loss, training with L2 loss, and training without either contrastive loss or L2 loss. As shown in Fig. 5, the original (left) audio features are clustered in adjacent frames, while the expression features do not have a clear distribution pattern. After training with L2 loss (middle), the audio features and expression features are simply pulled closer together. However, after training with contrastive learning loss (right), the expression features show a clustered aggregation phenomenon similar to the audio features, indicating that the audio information has influenced the distribution of expression features and led to the cluster of positive samples.

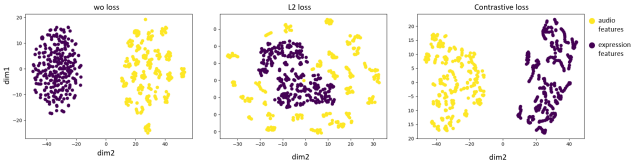

Figure 5: T-SNE plots of audio features and expression features extracted by the reconstruction module.

Comparison of Contractive Loss and L2 loss: We also conducted a comparative experiment to investigate the im-

provement in reconstruction results using contrastive loss. We replaced contrastive loss with L2 loss, and the results are shown in Tab. 4. The results demonstrate that compared to L2 loss, contrastive learning loss can better improve the reconstruction of the mouth region.

Metric	**L2 loss**	**Contrastive loss**
Lip Vertex Error ↓	$1.068e^{-2}$	$\mathbf{0.995}e^{-2}$

Table 4: Ablation study on pre-train and finetune.

4. Generalizability of Our Framework

To show that our proposed framework can be easily applied to other audio2mesh methods, we also conducted a verification based on VOCA [1]. We first pre-trained VOCA with pseudo-4D meshes reconstructed from the MEAD dataset and then fine-tuned it with the VOCAset. Due to the convolutional module in VOCA, we chose only 8 subjects from the MEAD dataset (42 subjects in total) with the mildest emotional intensity in the pre-training phase. The results, as presented in Tab. 5, indicate that VOCA with our framework achieves a lower lip vertex error. This improvement is due to the larger amount of training data, which demonstrates that our framework effectively addresses the data scarcity issue of 4D scans. With more reconstructed meshes from 2D videos, we can expect even greater improvements in future works.

Metric	VOCA wo. pre.	VOCA w. pre.
Lip Vertex Error ↓	$6.428e^{-3}$	$\mathbf{6.108}e^{-3}$

Table 5: Comparison between pre-trained and non-pre-trained models of VOCA.

5. Description of the Supplementary Video

The supplementary video contains three parts:

1. In the first part, we show the speech-assisted 3D face reconstruction results on different video datasets along with other competitive monocular 3D face reconstruction methods (including 3DDFA-V2, Deep3DFace pytorch version, DECA, and EMOCA). From the videos, we can see that our method significantly outperforms other methods, especially on the lower face region, which plays a more important role in the downstream speech-driven facial animation task.

2. In the second part, we compare our speech-driven 3D facial animation method with FaceFormer (SOTA method in speech-driven facial animation), VOCA, and MeshTalk with different speeches. Besides, we also directly compare our animation results with the videos from FaceFormer's supplementary materials. It can be found that our animated faces are more expressive and perceptually reasonable, especially for some syllables that need pursed lips.

3. In the final part, we demonstrate the emotional controllability of our model. By pretraining the audio2mesh module on the reconstructed emotional talking head dataset (i.e., MEAD), our method can embed 7 emotions in the animated talking head.

References

[1] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019. 3

[2] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *CVPR*, pages 20311–20322, 2022. 2

[3] Panagiotis P Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Visual speech-aware perceptual 3d facial expression reconstruction from videos. *arXiv preprint arXiv:2207.11094*, 2022. 1