– Supplementary Material –
EmoTalk: Speech-Driven Emotional Disentanglement for 3D Face Animation

Ziqiao Peng[1] Haoyu Wu[1] Zhenbo Song[2] Hao Xu[3,6] Xiangyu Zhu[4]
Jun He[1] Hongyan Liu[5*] Zhaoxin Fan[1,6*]
[1]Renmin University of China [2]Nanjing University of Science and Technology
[3]The Hong Kong University of Science and Technology [4]Chinese Academy of Sciences
[5]Tsinghua University [6]Psyche AI Inc.
{pengziqiao, wuhaoyu556, hejun, fanzhaoxin}@ruc.edu.cn songzb@njust.edu.cn
hxubl@connect.ust.hk xiangyu.zhu@nlpr.ia.ac.cn liuhy@sem.tsinghua.edu.cn

In this supplementary material, we provide more details about EmoTalk, which consists of five parts: 1) The implementation details of EmoTalk, including the model architecture and parameter details; 2) The transform module from blendshape to FLAME head, including the transform method and calculation formula; 3) The comparison method with baselines, including the comparison objects and evaluation details; 4) The construction details of the 3D-ETF dataset, including data collection, preprocessing, and post-processing; 5) The implementation details of blendshape capture method.

1. Implementation details

EmoTalk's overall architecture is illustrated in Fig. 2 of the main paper. In order to improve the reproducibility and credibility of EmoTalk on the 3D emotional face animation generation task, we will further explain how we design and implement two key components: emotion disentangling encoder and emotion-guided feature fusion decoder.

1.1. Training details

The network receives preprocessed video and audio data as input. The video stream is converted to 30 frames per second, while the audio sampling rate is 16 kHz. A facial blendshape capture method generates facial parameters consisting of 52 blendshape coefficients per frame for the video data.

During the training process, the model is optimized end-to-end using the Adam optimizer [4]. The learning rate and batch size are set to $1e$ - 4 and 8, respectively. The model is trained on a single NVIDIA V100, and the entire network takes approximately 8 hours (80 epochs) to train.

*corresponding authors

1.2. Emotion disentangling encoder

To perform emotion disentanglement, we first convert the input audio signal to a sampling rate of 16 KHz. Then we encode it using temporal convolutional network (TCN) to process sequential data with convolutional architecture. Next, we use a linear interpolation layer to adjust the length of the encoded representation according to the target audio signal. For instance, if we want to reconstruct $\boldsymbol{A}_{c1,e1}$ using $\boldsymbol{A}_{c1,e2}$ and $\boldsymbol{A}_{c2,e1}$ as inputs, then we need to interpolate them to have the same length as $\boldsymbol{A}_{c1,e1}$. After that, we decode the interpolated representation using 24 transformer[9] blocks. Each transformer block has a model dimension of 1024, an inner dimension of 4096, and 16 attention heads. Finally, we obtain two feature vectors of dimension 1024 each, representing content and emotional information in the output audio signal from pre-trained models. We use a cross-reconstruction constraint method to optimize model parameters during the training process, which we detail in Sec 3.1 of the main paper.

1.3. Emotion-guided feature fusion decoder

We first map the output of the features by the emotion feature extractor and the content feature extractor to 256-dimensional and 512-dimensional vectors, respectively. Then we add two one-hot embeddings for emotion level and personal style, each mapped to a 32-dimensional vector. The emotion level is a binary variable indicating high or low intensity, while the personal style is a multivariate variable representing 24 different speakers. We concatenate these four features to form an 832-dimensional feature vector. We also add a periodic position encoding[2] of the same dimension to this vector. Moreover, we use a fully connected layer to reduce the dimension of the output of the features by the emotion encoder from 1024 to 832 for subsequent emotion guidance. For biased multi-head

self-attention and emotion-guided multi-head attention, we use four heads and set the dimension to 832 for each transformer decoder block. The concatenated features serve as the input sequence for the decoder, while emotional features serve as the output sequence from the last encoder layer, thus achieving emotion guidance. Finally, we feed the forward layer's output into the audio-blendshape decoder, which is a fully connected layer that maps between 832 dimensions and 52 dimensions blendshape coefficients. Thus we obtain emotion-enhanced blendshape coefficients.

2. Blendshape to FLAME transform module

The Blendshape[5] to FLAME[6] transform module converts blendshapes, which is a way of deforming a mesh by interpolating between different shapes, to a FLAME head, which is a 3D head model that captures variations in identity, expression, head pose and gaze. This transform module enables our model to transfer facial expressions across different virtual characters quickly. To achieve this conversion, we collaborated with professional animators to create 52 semantically meaningful FLAME head templates (see Fig. 1). These templates allow us to obtain the facial deformation parameters corresponding to blendshape and mesh head. We use blend linear skinning to interpolate between these parameters. Because blendshape labels have semantic meanings, they can quickly transfer facial motions across different virtual characters.

Specifically, after obtaining the blendshape coefficients output by EmoTalk, we perform linear weighting on the corresponding parameters of 52 FLAME head templates to obtain the vertex parameters of 5023*3 dimensions. The formula is as follows:

$$V_{flame} = \sum_{i=1}^{52} \beta V_i \tag{1}$$

where $\boldsymbol{V}_{flame}$ is the final output of FLAME head vertex coordinates, $\boldsymbol{V}_i$ is the vertex coordinate of the $\boldsymbol{i}^{th}$ FLAME head template, and $\boldsymbol{\beta}$ is the blendshape coefficient vector output by EmoTalk.

3. Baseline methods

We conducted a comparative analysis of EmoTalk with three state-of-the-art approaches, namely VOCA[1], MeshTalk[8], and FaceFormer[2]. To facilitate a comprehensive evaluation, we employed two distinct datasets, namely the RAVDESS and HDTF, both of which are processed through our facial blendshape capturing technique to obtain the ground truth. For each frame in the datasets, we calculated the blendshape coefficients and mapped them to the corresponding vertex parameters of the FLAME model using the transform module. Furthermore, we retrained the models of the three existing approaches using RAVDESS , HDTF and 3D-ETF datasets to improve their performance.

For VOCASET, we used the pre-trained models provided by VOCA and FaceFormer and retrained the MeshTalk model to evaluate the vertex error of these three methods on the VOCA-Test. It is worth noting that due to the absence of blendshape coefficients in the official VOCASET dataset and the images containing marked faces incompatible with our blendshape capturing approach, we are unable to train our model on this dataset. Instead, we directly evaluated the EmoTalk model, trained on the HDTF dataset, on VOCA-Test.

During the evaluation, while the other three methods computed the error directly between the output vertices and the ground truth, we needed to use a transfer module to convert the EmoTalk output from blendshape coefficients to mesh vertices to ensure comparability with other methods in the same dimension and eliminate any differences between output formats.

4. Dataset construction details

In this study, we constructed a large 3D emotional talking face (3D-ETF) dataset, where facial blendshape is used as the supervisory signal to reconstruct reliable 3D faces from 2D images. The facial blendshape capturing method is fine-tuned by animators to create numerous 3D facial animations from the RAVDESS[7] and HDTF[10] datasets.

Specifically, 1440 videos from the RAVDESS dataset and 385 videos from the HDTF dataset are processed by converting them into 30 frames per second and capturing the facial blendshape for each frame. To enhance the quality of the dataset and reduce frame-to-frame jitter, a Savitzky-Golay filter with a window length of 5 and a polynomial order of 2 is applied to the output blendshape coefficients, which significantly improved the smoothness of facial animation. The RAVDESS dataset generated 159,702 frames of blendshape coefficients, which amounts to approximately 1.5 hours of video content. Meanwhile, the HDTF dataset generated 543,240 frames of blendshape coefficients, which equates to approximately 5 hours of video content. All the generated blendshape coefficients are converted into mesh vertices using the transform module and included in the dataset. A supplementary video will demonstrate the effectiveness of our dataset.

5. Blendshape capture method

Our sophisticated blendshape capture method predicts corresponding blendshape coefficients from input video streams using a neural network model, which is then manually fine-tuned by professional animators to achieve realistic facial reconstruction results that accurately capture human emotional expressions.

In this method, we use the "Live Link Face" application to collect a dataset consisting of images paired with corresponding blendshape data. The image preprocessing involved facial cropping and other necessary transformations before feeding them into a ResNet [3] architecture. The ResNet model was employed to produce 52 specific blendshape values as the output, and these values were constrained using the L2 loss function, ensuring precise regression of facial blendshapes.

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. 2

[2] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18780, 2022. 1, 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[4] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 1

[5] John P Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Frederic H Pighin, and Zhigang Deng. Practice and theory of blendshape facial models. *Eurographics (State of the Art Reports)*, 1(8):2, 2014. 2

[6] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 2

[7] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. 2

[8] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1153–1162. IEEE, 2021. 2

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[10] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 2

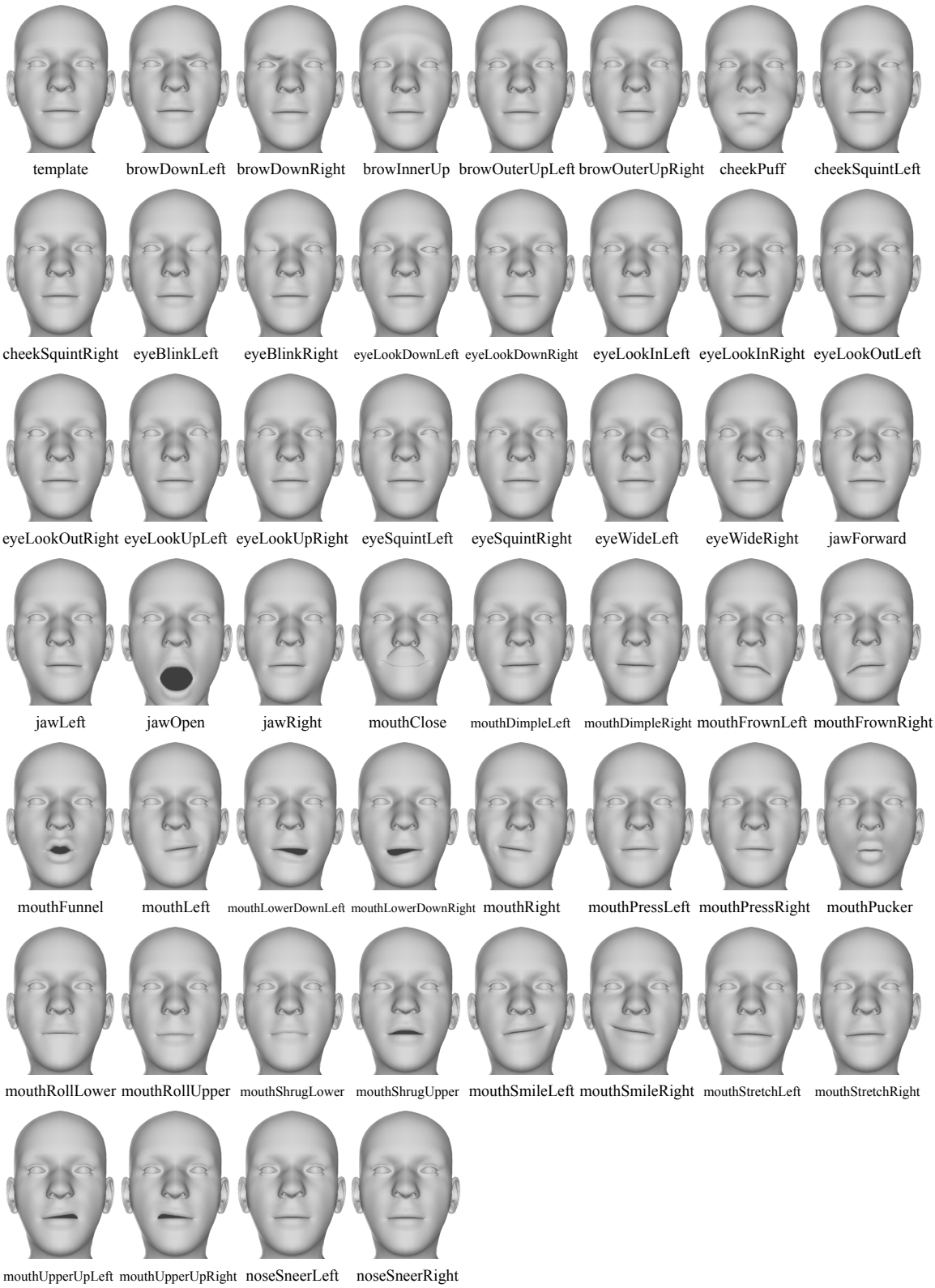

Figure 1. **Semantically Meaningful FLAME Head Templates.** We create 52 FLAME head templates that correspond to the blendshape coefficients, to achieve the transformation from the blendshape coefficients to the FLAME head model.