

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

Ziqiao Peng¹ Haoyu Wu¹ Zhenbo Song² Hao Xu^{3,6} Xiangyu Zhu⁴
 Jun He¹ Hongyan Liu^{5*} Zhaoxin Fan^{1,6*}

¹Renmin University of China ²Nanjing University of Science and Technology
³The Hong Kong University of Science and Technology ⁴Chinese Academy of Sciences
⁵Tsinghua University ⁶Psyche AI Inc.

{pengziqiao, wuhaoyu556, hejun, fanzhaoxin}@ruc.edu.cn songzb@njust.edu.cn
 hxubl@connect.ust.hk xiangyu.zhu@nlpr.ia.ac.cn liuh@sem.tsinghua.edu.cn

Abstract

Speech-driven 3D face animation aims to generate realistic facial expressions that match the speech content and emotion. However, existing methods often neglect emotional facial expressions or fail to disentangle them from speech content. To address this issue, this paper proposes an end-to-end neural network to disentangle different emotions in speech so as to generate rich 3D facial expressions. Specifically, we introduce the emotion disentangling encoder (EDE) to disentangle the emotion and content in the speech by cross-reconstructed speech signals with different emotion labels. Then an emotion-guided feature fusion decoder is employed to generate a 3D talking face with enhanced emotion. The decoder is driven by the disentangled identity, emotional, and content embeddings so as to generate controllable personal and emotional styles. Finally, considering the scarcity of the 3D emotional talking face data, we resort to the supervision of facial blendshapes, which enables the reconstruction of plausible 3D faces from 2D emotional data, and contribute a large-scale 3D emotional talking face dataset (3D-ETF) to train the network. Our experiments and user studies demonstrate that our approach outperforms state-of-the-art methods and exhibits more diverse facial movements. We recommend watching the supplementary video: <https://zqiaopeng.github.io/emotalk>

1. Introduction

Dynamic and realistic speech-driven facial animation has garnered growing interest in virtual reality [52, 11, 12], computer gaming [38, 10, 2], and film production [28, 53, 7]. For current commercial products, 3D face blendshape is handcrafted by animators, whereas manual scripts drive

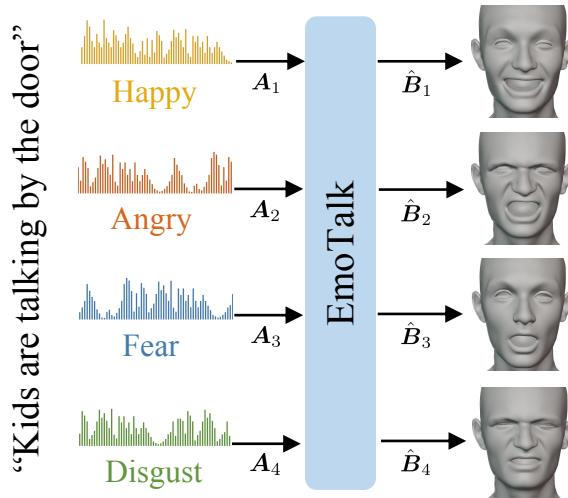


Figure 1. **Results of EmoTalk.** Given audio input expressing different emotions, EmoTalk produces realistic 3D facial animation sequences with corresponding emotional expressions as outputs.

facial expressions. Such a process demands substantial expenses and considerable time and labor. As deep learning techniques are utilized in various scenarios [39], deep end-to-end speech-driven facial animation [21, 8, 41, 9, 6] has been widely studied in industry and academia. Presently, learning-based 3D facial animations can not only produce high-quality animation effects but also facilitate cost reduction during production.

However, current methods mainly focus on improving the synchronization between lip movements and speech [44], neglecting the emotional variation of facial expressions. We argue that emotions are an essential aspect of human communication and expression, and emotion absence in 3D facial animations may cause the uncanny valley effect. It is a crucial issue to recover emotional expressions

*corresponding authors

for the speech-driven 3D face animation problem. In fact, emotional information is naturally contained in the speech, and extracting emotions is a crucial task for speech understanding [50]. Nevertheless, as audio content and emotion are entangled, it is hard to extract explicit content and emotion from a speech simultaneously. In order to generate rich emotional facial expressions, previous 2D facial animation methods encode the emotions manually and only learn the content feature from the speech [36, 4, 45]. By manipulating the emotion code, the facial decoder could achieve appropriate emotional modulation. Manually controlling may generate changeable emotions, but it could result in contradiction with the emotion in speech. For example, it does not conform to human intuition by inputting angry speech but outputting a happy expression.

To address this issue, we propose a novel speech-driven emotion-enhanced 3D facial animation method (Fig. 1) in this paper, where an emotion disentangling encoder and emotion-guided feature fusion decoder are proposed to consist of our key contribution, as illustrated in Fig. 2. For the emotion disentangling encoder, two distinct audio feature extractors [1] are introduced and utilized to extract two separate latent spaces for the content and emotion, respectively, which is exploited to decouple emotion and content. A cross-reconstruction loss is further presented to constrain the learning process to better disentangle the emotion and content from the speech. While for the emotion-guided feature fusion decoder, multiple different types of features are decoded by a Transformer [49] module with periodic positional encoding and emotion-guided multi-head attention, which will output 52 emotion-enhanced blendshape coefficients to represent the final human facial expressions. Extensive experiments show that our method significantly outperforms current state-of-the-art methods in terms of emotional expression by disentangling content and emotion.

To train the proposed network, emotional speeches with corresponding 3D facial expressions are required. However, as far as we know, there is no publicly available 3D emotional talking face dataset that we can use, posing a serious new challenge. To tackle the issue, a large-scale pseudo-3D emotional talking face dataset, termed the 3D-ETF dataset, is further introduced in our work. To build this dataset and make it more applicable, we first collaborated with several professional animators to create 52 FLAME head templates [27] that are semantic meaningful. Then, “pseudo” 3D blendshape labels are generated from images of large-scale audio-visual datasets [30, 58] by utilizing a well-established 3D facial blendshape capture system. Finally, the 3D-ETF dataset with both blendshape coefficients [26] and mesh vertices are constructed through blend linear skinning. Since its blendshape labels are semantic meaningful, the 3D-ETF dataset is versatile, allowing the facile transfer of facial movements among different virtual characters [37].

In summary, the main contributions of our work are as follows:

- We propose an end-to-end neural network for speech-driven emotion-enhanced 3D facial animation, which achieves various emotional expressions and outperforms existing state-of-the-art methods.
- We introduce the emotion disentangling encoder, which disentangles the emotion and content in the speech and makes the facial animation aware of clear emotional information.
- We present a large-scale 3D emotional talking face (3D-ETF) dataset including both blendshape coefficients and mesh vertices. We have implemented parameterized transformations for blendshape coefficients and the FLAME model, allowing for efficient conversion between various facial animations.

2. Related Work

2.1. Speech-driven 3D facial animation

Previously, numerous studies have been conducted on 2D talking head generation [5, 55, 32, 13, 56, 46, 20, 15], which uses image-driven or speech-driven approaches to create realistic videos of speaking individuals. However, these methods are not applicable to 3D character models that are widely used in 3D games and virtual reality interactions. Therefore, speech-driven 3D facial animation has attracted more attention recently [3, 21, 8, 47, 18, 41, 29, 9, 6].

One of the challenges in this field is the lack of high-quality datasets for various emotions. For example, VOCA [8] uses time convolutions and control parameters to generate realistic character animation from any speech signal and a static character mesh. Still, it only produces decent mouth movements due to the limited upper face movement in the VOCASET dataset [8]. Similarly, FaceFormer [9] uses a Transformer-based model to obtain contextually relevant audio information and generates continuous facial movements in an autoregressive manner. It improves multi-source generalization and achieves more precise changes in mouth movements than VOCA but still does not enhance facial expressions because it also uses the VOCASET dataset.

MeshTalk [41] focuses on the upper part of the face, which is lacking in VOCA, and creates a categorical latent space for facial animation, disentangles audio-correlated and audio-uncorrelated movements through cross-modality loss, thereby synthesizing audio-uncorrelated movements such as blinking and eyebrow movements. Although MeshTalk has achieved upper face movement, the current methods still need to solve the problem of lack of emotion in 3D facial animation due to the absence of emotional facial animation datasets.

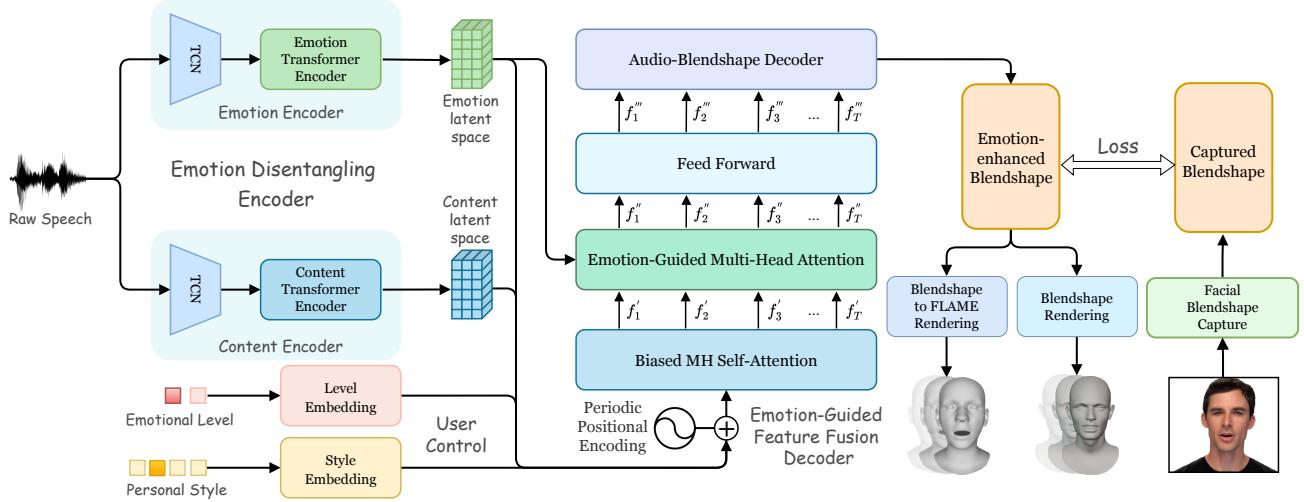


Figure 2. **Overview of EmoTalk.** Given a speech input $\mathbf{A}_{1:T}$, emotional level \mathbf{l} , and personal style \mathbf{p} as inputs, our model disentangles the emotion and content in the speech using two latent spaces. The features extracted from these latent spaces are combined and fed into the emotion-guided feature fusion decoder, which outputs emotion-enhanced blendshape coefficients. These coefficients can be used to animate a FLAME model or rendered as an image sequence.

2.2. Speech emotion recognition and disentanglement

Speech emotion recognition (SER) is an essential but challenging task for generating realistic talking head animations. Various techniques have been proposed in the paper to extract emotions from speech signals, such as traditional speech analysis and classification methods [35, 42, 23, 17, 43, 16, 19]. In this paper, we focus on deep learning-based SER techniques to learn features from speech signals. For example, Mekruksavanich *et al.* [31] used one-dimensional CNNs [25] to achieve 96.60% accuracy in classifying negative emotions from a Thai language dataset. Yenigalla *et al.* [54] combined phoneme sequences and spectrograms as inputs to a CNN model and obtained better SER performance than using either input alone.

One of the key challenges in speech processing is to separate emotion from content, which enables the neural network to learn more specific features. This process is called emotion disentanglement. Several methods exist for achieving emotion disentanglement and reconstruction in speech using different techniques, such as variational autoencoding Wasserstein generative adversarial networks (VAW-GANs) [59], cross-speaker emotion transfer [57], and Mel Frequency Cepstral Coefficient (MFCC) [33] with Dynamic Time Warping (DTW) [34]. In this paper, we build on the work of Ji *et al.* [20], who decomposed speech signals into two decoupled spaces and fed them into a facial synthesis module. We propose an improved emotion disentangling encoder for 3D facial animation generation, which will be described in detail in Sec. 3.1.

3. Method

We propose a 3D facial animation model that can reconstruct facial expressions with rich emotions from speech signals, enabling users to control emotional level and personal style. Let $\mathbf{A}_{1:T} = (\mathbf{a}_1, \dots, \mathbf{a}_T)$ be a sequence of speech snippets, and each $\mathbf{a}_t \in \mathbb{R}^D$ has D samples to align to the corresponding (visual) frame \mathbf{b}_t . Let $\mathbf{B}_{1:T} = (\mathbf{b}_1, \dots, \mathbf{b}_T), \mathbf{b}_t \in \mathbb{R}^{52}$ be a T -length sequence of face blendshape coefficients, and each frame is represented by 52 values. The whole pipeline of our approach is revealed in Fig. 2. By analyzing emotional information from any arbitrary speech signal $\mathbf{A}_{1:T}$, our method is capable of producing differentiated face coefficients $\hat{\mathbf{B}}_{1:T}$. Moreover, the proposed model takes a user-controllable emotional level $\mathbf{l} \in \mathbb{R}^2$ as input, which allows users to modulate the strength of the expressed emotions in the resulting facial animations. Personal style $\mathbf{p} \in \mathbb{R}^{24}$ inputs can also be manipulated by users to have different speaking habits. These two parameters are the same one-hot encoding as [48]. Then, the decoder predicts facial coefficients $\hat{\mathbf{B}}_{1:T} = (\hat{\mathbf{b}}_1, \dots, \hat{\mathbf{b}}_T)$ conditioned on speech representations $\mathbf{A}_{1:T}$, the emotional level \mathbf{l} , and the personal style \mathbf{p} . Formally,

$$\hat{\mathbf{b}}_t = \text{EmoTalk}_\theta(\mathbf{a}_t, \mathbf{l}, \mathbf{p}), \quad (1)$$

where θ indicates the model parameters. For the convenience of describing detailed network components, let $\mathbf{A}_{ci,ej}$ denote the sample data pertaining to the i^{th} content and j^{th} emotion in the audio sample, whereas $\mathbf{B}_{ci,ej}$ denote the sample data pertaining to the i^{th} content and j^{th} emotion in the blendshape coefficients sample. Both rep-

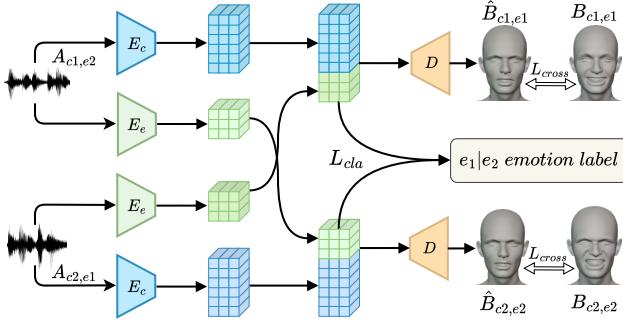


Figure 3. Emotion Disentangling Encoder. Various inputs of speech, conveying different contents and emotions, are processed to generate cross-reconstructed blendshape coefficients representing distinct combinations of facial expressions.

resentations will be employed in the following sections to introduce the details of our method.

3.1. Emotion disentangling encoder

The intricate relationship between speech and facial expressions makes it arduous to learn the mapping from speech to emotional facial expressions directly. To address this issue, we propose an improved emotion disentangling encoder for 3D facial animation generation, inspired by Ji *et al.* [20]. To the best of our knowledge, this is the first work that applies emotion disentanglement to this task. Our module simplifies and enhances the original disentanglement module in several ways. First, we replace the MFCC [33] feature extractor, which cannot capture rich speech information and has a complex input process, with a pre-trained audio feature extractor wav2vec 2.0 [1]. Second, we streamline the disentanglement process to enhance its conciseness and comprehensibility. Third, we transform the module into an end-to-end form that directly outputs 52 blendshape coefficients required for facial animation, allowing the model to receive better constraints during training.

Reorganization and disentanglement. As illustrated in Fig. 3, the emotion disentangling encoder is designed to disentangle short-term content features from long-term emotion features in speech.

Nevertheless, the module cannot guarantee the disentanglement between content and emotion. To achieve this objective, we utilize pseudo-training pairs that combine diverse emotions and contents as input and require the network to reconstruct the corresponding ground truth samples as output. This approach compels the network to acquire disentangled content and emotion representations, which can better capture both aspects of speech and enhance the overall performance of the model.

To separate content and emotion features in speech, two pre-trained audio models [1] are used as feature extractors E_c and $E_e \in \mathbb{R}^{1024}$, which are fine-tuned on content

and emotion, respectively. The pre-trained models’ temporal convolutional network(TCN) layer [24] is fixed during fine-tuning since it is trained on a considerable amount of audio data. We input two audios $A_{c1,e2}$ and $A_{c2,e1}$, where the subscript c denotes text content, and e denotes audio emotion. Content features $c1$ and $c2$ are extracted using $E_c(A_{c1,e2})$ and $E_c(A_{c2,e1})$, respectively. Emotion features $e1$ and $e2$ are extracted using $E_e(A_{c2,e1})$ and $E_e(A_{c1,e2})$, respectively. The content and emotion features are concatenated and fed into a decoder module that outputs face blendshape coefficients for reconstruction. Pseudo-training pairs comprising different combinations of content and emotion are used as input, and the network is required to reconstruct the corresponding ground truth samples as output, namely $\hat{B}_{c1,e1}$ and $\hat{B}_{c2,e2}$, for constraints with real samples $B_{c1,e1}$ and $B_{c2,e2}$. This approach enforces disentanglement between content and emotion features by requiring that they can be combined to reproduce both aspects of speech.

3.2. Emotion-guided feature fusion decoder

In this work, we propose an emotion-guided feature fusion decoder that maps audio to 3D facial animation coefficients using emotional information from audio. This approach aims to generate more expressive facial animations. This module consists of four components: emotion features $F_e \in \mathbb{R}^{256}$ and content features $F_c \in \mathbb{R}^{512}$ extracted from two latent spaces, personal style features $F_p \in \mathbb{R}^{32}$ that control the individual characteristics of facial expressions, and emotional level features $F_l \in \mathbb{R}^{32}$ that regulate the degree of emotional expression. These four features are concatenated along the same dimension and subsequently inputted into the emotion-guided feature fusion decoder.

To generate the 3D blendshape coefficients from the fused feature, we employ a module similar to the Transformer [49] decoder. The input feature F is first encoded with periodic positional encoding [9], which captures the stable open and close times of lip movements during speech. Then, a biased multi-head self-attention layer that integrates positional encoding into multi-head attention layers inspired by attention with linear biases (ALiBi) [40] produces f'_t , which assigns higher weights to closer information in the mask layer and focuses on the changes between adjacent actions. Subsequently, an emotion-guided multi-head attention that combines f'_t and the output $E_e(A_{ci,ej})$ of the emotion latent space is proposed. This module enhances the emotional expressiveness of 3D animated faces, as demonstrated by experiments conducted in this study (Tab. 5). Finally, f''_t is fed into a feed-forward layer that outputs f'''_t , which is then passed through an audio-blendshape decoder implemented as a fully connected layer that outputs 52 blendshape coefficients.

3.3. Loss function

To train our neural network, we employ a loss function that comprises four distinct components: cross-reconstruction loss, self-reconstruction loss, velocity loss, and classification loss. The overall function is given by:

$$L = \lambda_1 L_{cross} + \lambda_2 L_{self} + \lambda_3 L_{vel} + \lambda_4 L_{cls}, \quad (2)$$

where $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, $\lambda_3 = 0.5$ and $\lambda_4 = 0.1$ in all of our experiments. We provide a detailed explanation of each of these components below.

Cross-reconstruction loss. In order to disentangle emotional content from speech signals, as described in Sec. 3.1, we train our network to reconstruct various cross combinations and generate new blendshape coefficients. Given input audio $A_{c1,e2}$ and $A_{c2,e1}$, the encoder decomposes them and then reconstructs new combinations, which are compared with the ground truth blendshape coefficients $B_{c1,e1}$ and $B_{c2,e2}$. The formula is as follows:

$$\begin{aligned} L_{cross} = & \|D(E_c(A_{c1,e2}), E_e(A_{c2,e1})) - B_{c1,e1}\|^2 \\ & + \|D(E_c(A_{c2,e1}), E_e(A_{c1,e2})) - B_{c2,e2}\|^2, \end{aligned} \quad (3)$$

where D is the emotion-guided feature fusion decoder for reconstructing the cross combinations.

Self-reconstruction loss. While constraining the quality of the reconstructed output using cross-reconstruction, we also require the network to reconstruct its ground truth blendshape coefficients. The self-reconstruction loss can be expressed as:

$$L_{self} = \|D(E_c(A_{c1,e2}), E_e(A_{c1,e2})) - B_{c1,e2}\|^2. \quad (4)$$

Velocity loss. To address the issue of jittery output frames when using only reconstruction loss, we introduce a velocity loss to induce temporal stability, which considers the smoothness of prediction and ground truth in the sequence context. By incorporating this loss, our model is encouraged to produce smoother and more realistic facial expressions. The velocity loss can be expressed as:

$$L_{vel} = \left\| (\hat{b}_t - \hat{b}_{t-1}) - (b_t - b_{t-1}) \right\|^2, \quad (5)$$

Classification loss. Due to the inherent difficulty of explicitly discerning the separability of emotional latent space during the disentangling process, we introduce a classification loss to supervise the output of the emotion extractor E_e and enhance its ability to discriminate between different emotions. The classification loss is defined as:

$$L_{cls} = - \sum_i \sum_{c=1}^M (\mathbf{y}_{ic} * \log \mathbf{p}_{ic}), \quad (6)$$

where M represents the number of distinct emotion categories, \mathbf{y}_{ic} is the observation function that determines whether sample i carries the emotion label c , and \mathbf{p}_{ic} denotes the predicted probability that sample i belongs to class c .

3.4. Datasets construction

Due to the scarcity of 3D talking face data with emotions, no such data is publicly available. To acquire such data, professional equipment and actors who can utter the same sentence with varied emotions are required, which entails high expenses. However, numerous 2D emotional audio-visual datasets exist. We employ facial blendshapes as a supervisory signal, which facilitates the reconstruction of plausible 3D faces from 2D images. Then, We extract blendshape coefficients from two datasets using a sophisticated blendshape capture method¹ which result accurately capture human emotional expressions (Fig. 4). A large 3D emotional talking face (3D-ETF) dataset consisting of approximately 700,000 frames of blendshape coefficients, spanning over 6.5 hours, is constructed using this method. Through blend linear skinning, both blendshape coefficients [26] and mesh vertices are built for the 3D-ETF dataset, filling a gap in 3D facial animation datasets, especially regarding emotional expression data and providing vivid and lifelike human facial expressions.

4. Experiments

4.1. Datasets

Two widely used 2D audio-visual datasets were utilized to construct the 3D-ETF dataset: RAVDESS [30] and HDTF [58].

The RAVDESS dataset [30], also known as the Ryerson Audio-Visual Database of Emotional Speech and Song, is a multi-modal emotion recognition dataset comprising 24 actors (12 male, 12 female) and 1440 video clips of short speeches. The dataset was captured with high-quality audio and video recordings, and the actors were instructed to express specific emotions, including neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. A random selection of 80% of the dataset was used for training, 10% for validation, and 10% for testing.

The High-Definition Talking Face (HDTF) dataset [58] is a collection of approximately 16 hours of 720P-1080P videos sourced from YouTube over the past few years. The dataset includes over 300 subjects and 10k different sentences. Five hours of videos from the HDTF dataset were selected for mouth shape generalization and then partitioned into training, validation, and testing sets in the same proportion as the RAVDESS dataset.

¹Details can be found in the Supplementary Material.

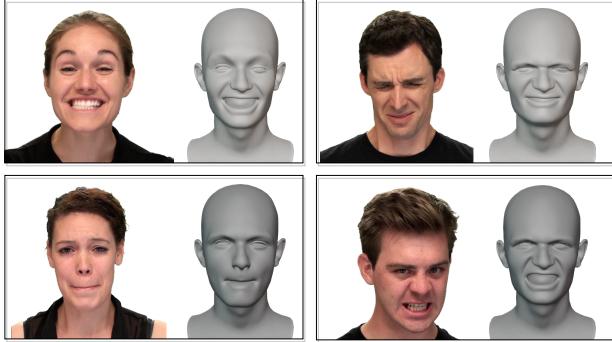


Figure 4. Facial Blendshape Capture. Input video streams of different expressions, and outputs the blendshape coefficients of the corresponding expressions.

	RAVDESS (emotion)	HDTF (no emotion)		
Method	LVE(mm)↓	EVE(mm)↓	LVE(mm)↓	EVE(mm)↓
VOCA [8]	5.091	4.188	4.447	3.286
MeshTalk [41]	3.459	3.386	3.886	3.124
FaceFormer [9]	3.247	3.757	3.374	3.142
Ours	2.762	2.493	2.892	2.364

Table 1. Quantitative evaluation results on RAVDESS and HDTF datasets. The lip vertex error (LVE) and emotional vertex error (EVE) of our method are lower than those of the current state-of-the-art methods.

4.2. Quantitative evaluation

To measure lip synchronization, we calculated the lip vertex error (LVE) as used in MeshTalk [41] and FaceFormer [9]. This evaluation metric computes the average ℓ_2 error of the lips in the test set. For a single frame, LVE is defined as the maximum ℓ_2 error among all lip vertices. Since LVE alone cannot reflect the full emotional expression, we proposed an emotional vertex error (EVE). To compute EVE, vertex indexes in the eye and forehead regions on the FLAME template are first selected. Similar to LVE, the EVE measures the maximum ℓ_2 error of the vertex coordinate displacement in the interested region, and the average LVE over the test set is reported as the evaluation metric.

We retrained VOCA, MeshTalk, FaceFormer, and our method (EmoTalk) on the RAVDESS and HDTF datasets. The blendshape coefficients were converted into mesh vertices (5023*3) corresponding to the FLAME model, which was used as ground truth. Tab. 1 shows LVE and EVE evaluation results. EmoTalk achieved lower lip error and emotion expression error than the three previous methods. The proposed model has more accurate lip movements and better emotional expression.

Generalization analysis. The model could not be trained

Method	LVE(mm)↓	Train on VOCASET
VOCA [8]	4.704	✓
MeshTalk [41]	4.513	✓
FaceFormer [9]	4.418	✓
Ours	4.134	✗

Table 2. Quantitative evaluation results on VOCA-Test. Our method exhibits strong generalization capability in zero-shot cases and outperforms the current state-of-the-art methods.

Method	RAVDESS	HDTF	MEAD	VOCASET
VOCA	2.700	2.427	2.236	2.292
MeshTalk	2.139	1.868	2.058	2.070
FaceFormer	1.958	1.391	1.852	1.944
Ours	1.648	0.626	1.498	1.914 (zero-shot)

Table 3. Quantitative evaluation results of lip average ℓ_2 error.

on VOCASET because access to the corresponding blendshape coefficients was unavailable, and the blendshape capture method was incompatible with the marked facial images provided by the official dataset. Nevertheless, we evaluated our model on this dataset by converting the output blendshape coefficients to mesh vertices and comparing them with ground truth.

As reported in Tab. 2, EmoTalk outperformed the other methods on VOCA-Test, even in zero-shot settings. This could be attributed to two reasons: (i) controlling facial animation through blendshape coefficients has a higher generalization ability than predicting vertex offsets based on mesh, and (ii) sufficient 2D datasets can also enable the model to learn complex relations between speech and facial expressions, thus achieving better results.

Robustness analysis. When using the lip maximum ℓ_2 error metric, there may be a potential impact of outliers present in the dataset. To mitigate the impact of outliers and present a more comprehensive evaluation, we additionally computed the lip average ℓ_2 error for proposed method and introduced the MEAD dataset [51]. In Tab. 3, we present the results of the lip average ℓ_2 errors obtained from our method and previous methods. The analysis of these errors demonstrates the superior performance of our proposed method over multiple datasets, substantiating its effectiveness and robustness in comparison to existing methods.

4.3. Qualitative evaluation

As audio and facial movements cannot be evaluated solely based on indicators and require human perceptual evaluation, we conducted a qualitative assessment of our model from two perspectives.

Lip synchronization. We compared our model with VOCA

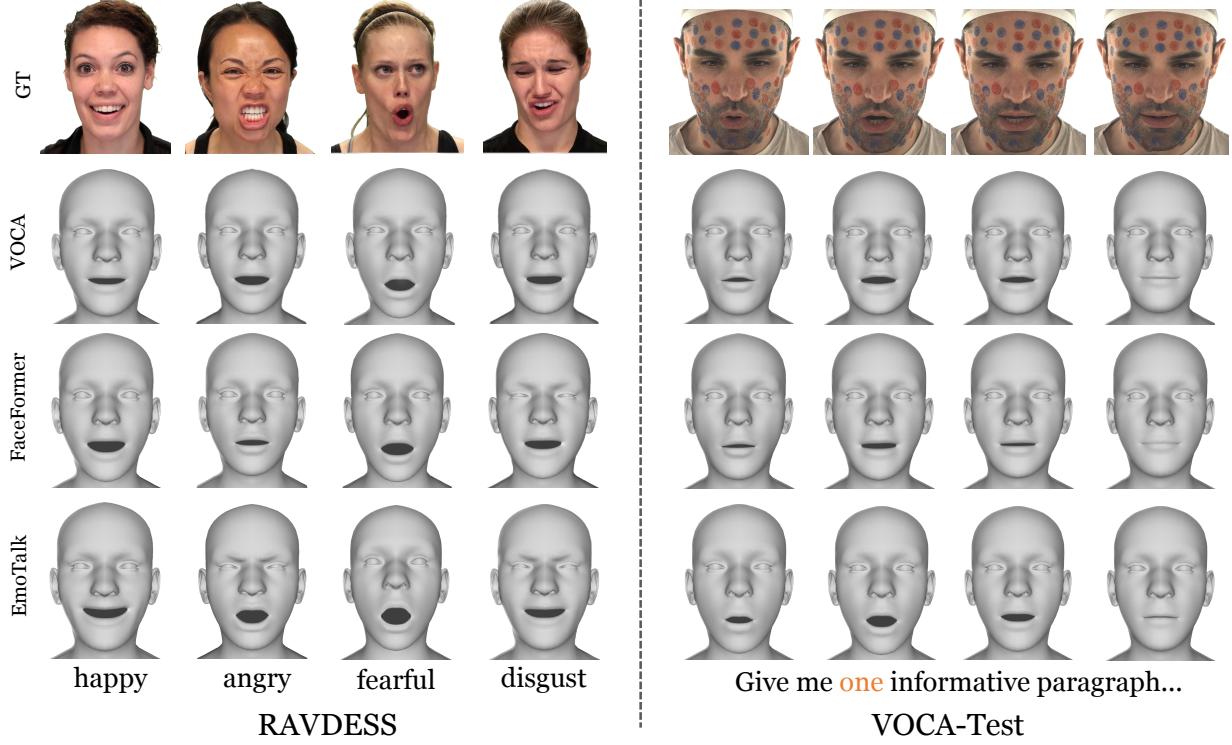


Figure 5. Qualitative comparison of facial movement by different methods on RAVDESS (left) and VOCA-Test (right). On RAVDESS, we generate facial animations of saying the word “dog” with different emotions. Our method can produce expressive facial movements that match the emotions. On VOCA-Test, we generate facial animations of saying the word “one” without emotion. Our method can achieve similar performance to the ground truth, and the range of motion is noticeable.

and FaceFormer by feeding them identical audio inputs and generating corresponding facial animations. The results showed that the proposed model exhibited more pronounced lip movements and better alignment with human speech patterns. Even in cases of rapid mouth movements, such as pronouncing the word “shy”, where the lips should gradually open and then close, the proposed model captured this lip synchronization more effectively (Fig. 5). Moreover, the proposed model’s ability to close the mouth was comparable to the results trained on high-precision scanned facial datasets.

Emotional expression. Previous methods were not optimized for emotional expression, resulting in limited facial expressions for different speech patterns. However, our method could clearly demonstrate variations among emotions. For example, in anger emotion, the movements of raising and lowering eyebrows could distinctly reflect the emotional information conveyed by speech (Fig. 5). A supplementary video provides more detailed comparisons.

4.4. User study

To evaluate the proposed model more thoroughly, We designed a comprehensive user questionnaire like [41] and performed a comparative analysis with MeshTalk [41] and FaceFormer [9] using the FLAME template. Since

our model incorporates emotional components, we devised three sub-tasks: full-face comparison, lip synchronization comparison (by covering the area above the nose), and emotion expression comparison (by covering the area below the nose). We selected twenty sentences from the RAVDESS and VOCASET test datasets as our test cases, ultimately formulating 120 multiple-choice questions. The questionnaire system randomly presented a pair of comparison videos to users, allowing them to choose which video is more realistic. We counted how many users chose our result versus the competitor’s result, and the ratio of user choice was calculated for satisfaction evaluation.

Specifically, compared with MeshTalk and FaceFormer, our model received the most positive feedback from participants, surpassing MeshTalk and FaceFormer in full-face voting by 65.9% and 64.6%, respectively. Notably, in terms of emotional expression, our model displays a huge advantage over alternative approaches. Overall, most participants considered our method superior to MeshTalk and FaceFormer. Detailed user choice results are shown in Tab. 4.

4.5. Ablation experiment

We conducted an ablation study to examine the contributions of different components of our model. The essential modules, loss functions, and datasets were individually

Method	Competitors	Ours
Ours vs. MeshTalk		
full-face	34.1%	65.9%
lip sync	38.7%	61.3%
emotion expression	31.5%	68.5%
Ours vs. FaceFormer		
full-face	35.4%	64.6%
lip sync	40.9%	59.1%
emotion expression	30.8%	69.2%

Table 4. **User study results.** We devised three sub-tasks, namely full-face comparison, lip synchronization comparison, and emotion expression comparison.

studied to examine their effects on the evaluation metrics. In Tab. 5, it is shown that a significant increase in emotional expression error ensued when removing the emotion disentangling encoder, which demonstrated the critical role of EDE in emotional learning and expression. Similarly, removing the emotion-guided multi-head attention also witnessed a certain increase in EVE, which indicated the effectiveness of the emotion guidance module in enhancing emotional expression.

From the loss function aspect, removing velocity loss leads to a slight drop in performance, but it caused noticeable jitter in the output of facial animation. Removing classification loss clearly increased the EVE, suggesting that the feature extractor could distinguish emotions less effectively. Then we train our model without using the HDTF dataset to investigate the LVE changes. It is observed that the LVE increases by about $0.5mm$, and the performance of lip vertex prediction decayed dramatically. This indicates that training with the HDTF dataset is able to learn more about mapping relationships between lip movements and speech. Finally, we replaced our emotion disentangling encoder with the method proposed by Ji et al [20], because they used MFCC as the audio feature extractor, and their approach required staged training during the processing. As a result, we observed an increase in errors during the evaluations of LVE and EVE. This indicates the effectiveness of the improvements we made on the Ji et al.’s method [20].

5. Limitations

Our method still has some limitations we plan to address in future work. First, our method relies on a large-scale audio pre-training model, which increases the inference time and hinders real-time applications. Second, our network outputs 52 blendshape coefficients, which do not include head movements, *e.g.* head shakes and rotations. A possible solution is to combine blendshape coefficients with the FLAME model [27] to control both facial expressions and head movements. Third, our training data is derived from

	LVE (mm)	EVE (mm)
Ours	2.762	2.493
w/o Emotion Disentangling Encoder	3.126	3.076
w/o Emotion-Guided Multi-Head Attention	2.907	2.832
w/o L_{vel} Loss	2.813	2.775
w/o L_{cls} Loss	3.096	2.815
w/o HDTF Dataset replace our Encoder with Ji et al.’s [20]	3.254 3.583	2.806 2.973

Table 5. **Ablation study for our components.** We show the LVE and EVE in different cases.

2D images. The pseudo-3D data is not as precise as 3D-scanned data and thus cannot represent the skin’s micro facial expressions. As a result, our method can only reflect the overall emotional state of the animated face. We intend to collect more emotional data using professional instruments in the future and share it with the research community.

6. Conclusion

This paper proposes a novel method for generating speech-driven 3D face animation that effectively conveys emotions. Our method consists of two key components: an emotion disentangling encoder and an emotion-guided feature fusion decoder. The former segregates the speech into its emotional and content components, providing clear emotional information for facial animation. The latter enhances the expressiveness of facial animation by emphasizing emotion-related features. To address the problem of missing 3D emotional talking face data, we construct a large-scale 3D emotional talking face (3D-ETF) dataset that contains blendshape coefficients and mesh vertices. Additionally, we have implemented parameterized transformations for blendshape coefficients and the FLAME model, allowing for efficient conversion between various facial animations. Experimental results demonstrate that our method outperforms existing state-of-the-art methods and receives better user experience feedback. Our work contributes to virtual reality applications. It can enable more realistic and immersive virtual experiences with emotional talking faces.

Acknowledgments

This work was supported in part by National Key Research and Development Program of China under Grant No. 2020YFB2104101 and National Natural Science Foundation of China (NSFC) under Grant Nos. 62172421, 62072459 and 71771131.

References

- [1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. [2](#), [4](#)
- [2] Elizabeth A Boyle, Ewan W MacArthur, Thomas M Connolly, Thomas Hainey, Madalina Manea, Anne Kärki, and Peter Van Rosmalen. A narrative literature review of games, animations and simulations to teach research methods and statistics. *Computers & Education*, 74:1–14, 2014. [1](#)
- [3] Yong Cao, Wen C Tien, Petros Faloutsos, and Frédéric Pighin. Expressive speech-driven facial animation. *ACM Transactions on Graphics (TOG)*, 24(4):1283–1302, 2005. [2](#)
- [4] Che-Jui Chang, Long Zhao, Sen Zhang, and Mubbasir Kapadia. Disentangling audio content and emotion with adaptive instance normalization for expressive facial animation synthesis. *Computer Animation and Virtual Worlds*, 33(3–4):e2076, 2022. [2](#)
- [5] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7832–7841, 2019. [2](#)
- [6] Liyang Chen, Zhiyong Wu, Jun Ling, Runnan Li, Xu Tan, and Sheng Zhao. Transformer-s2a: Robust and efficient speech-to-animation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7247–7251. IEEE, 2022. [1](#), [2](#)
- [7] Sixiang Chen, Tian Ye, Yun Liu, Taodong Liao, Jingxia Jiang, Erkang Chen, and Peng Chen. Msp-former: Multi-scale projection transformer for single image desnowing. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. [1](#)
- [8] 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. [1](#), [2](#), [6](#), [13](#)
- [9] 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](#), [4](#), [6](#), [7](#), [12](#), [13](#)
- [10] Zhaoxin Fan, Yulin He, Zhicheng Wang, Kejian Wu, Hongyan Liu, and Jun He. Reconstruction-aware prior distillation for semi-supervised point cloud completion. *arXiv preprint arXiv:2204.09186*, 2022. [1](#)
- [11] Zhaoxin Fan, Zhenbo Song, Jian Xu, Zhicheng Wang, Kejian Wu, Hongyan Liu, and Jun He. Object level depth reconstruction for category level 6d object pose estimation from monocular rgb image. In *European Conference on Computer Vision*, pages 220–236. Springer, 2022. [1](#)
- [12] Zhaoxin Fan, Yazhi Zhu, Yulin He, Qi Sun, Hongyan Liu, and Jun He. Deep learning on monocular object pose detection and tracking: A comprehensive overview. *ACM Computing Surveys*, 55(4):1–40, 2022. [1](#)
- [13] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021. [2](#)
- [14] 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. [13](#)
- [15] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3397–3406, 2022. [2](#)
- [16] M Shamim Hossain and Ghulam Muhammad. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, 49:69–78, 2019. [3](#)
- [17] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. Speech emotion recognition using cnn. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 801–804, 2014. [3](#)
- [18] Ahmed Hussen Abdelaziz, Barry-John Theobald, Paul Dixon, Reinhard Knothe, Nicholas Apostoloff, and Sachin Kajareker. Modality dropout for improved performance-driven talking faces. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 378–386, 2020. [2](#)
- [19] Manas Jain, Shruthi Narayan, Pratibha Balaji, Abhijit Bhowmick, Rajesh Kumar Muthu, et al. Speech emotion recognition using support vector machine. *arXiv preprint arXiv:2002.07590*, 2020. [3](#)
- [20] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089, 2021. [2](#), [3](#), [4](#), [8](#)
- [21] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. [1](#), [2](#)
- [22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. [12](#)
- [23] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. Emotion recognition by speech signals. In *Eighth European conference on speech communication and technology*, 2003. [3](#)
- [24] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks: A unified approach to action segmentation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part III 14*, pages 47–54. Springer, 2016. [4](#)
- [25] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. [3](#)
- [26] 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, 5, 12
- [27] 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, 8, 12
- [28] Chen Liu. An analysis of the current and future state of 3d facial animation techniques and systems. 2009. 1
- [29] Jingying Liu, Binyuan Hui, Kun Li, Yunke Liu, Yu-Kun Lai, Yuxiang Zhang, Yebin Liu, and Jingyu Yang. Geometry-guided dense perspective network for speech-driven facial animation. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4873–4886, 2021. 2
- [30] 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, 5, 13
- [31] Sakorn Mekruksavanich, Anuchit Jitpattanakul, and Narit Hnoohom. Negative emotion recognition using deep learning for thai language. In *2020 joint international conference on digital arts, media and technology with ECTI northern section conference on electrical, electronics, computer and telecommunications engineering (ECTI DAMT & NCON)*, pages 71–74. IEEE, 2020. 3
- [32] Gaurav Mittal and Baoyuan Wang. Animating face using disentangled audio representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3290–3298, 2020. 2
- [33] Lindasalwa Muda, Mumtaj Begam, and Irraivan Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083*, 2010. 3, 4
- [34] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007. 3
- [35] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623, 2003. 3
- [36] Foivos Paraperas Papantoniou, Panagiotis P Filntisis, Petros Maragos, and Anastasios Roussos. Neural emotion director: Speech-preserving semantic control of facial expressions in “in-the-wild” videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18781–18790, 2022. 2
- [37] Chandan Pawaskar, Wan-Chun Ma, Kieran Carnegie, John P Lewis, and Taehyun Rhee. Expression transfer: A system to build 3d blend shapes for facial animation. In *2013 28th International Conference on Image and Vision Computing New Zealand (IVCNZ 2013)*, pages 154–159. IEEE, 2013. 2
- [38] Heng Yu Ping, Lili Nurliyana Abdullah, Puteri Suhaiza Sulaiman, and Alfian Abdul Halin. Computer facial animation: A review. *International Journal of Computer Theory and Engineering*, 5(4):658, 2013. 1
- [39] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and Sundaraja S Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, 51(5):1–36, 2018. 1
- [40] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021. 4
- [41] 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. 1, 2, 6, 7, 13
- [42] Björn Schuller, Gerhard Rigoll, and Manfred Lang. Hidden markov model-based speech emotion recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)., volume 2*, pages II–1. Ieee, 2003. 3
- [43] Björn W Schuller. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99, 2018. 3
- [44] Changchong Sheng, Gangyao Kuang, Liang Bai, Chemping Hou, Yulan Guo, Xin Xu, Matti Pietikäinen, and Li Liu. Deep learning for visual speech analysis: A survey. *arXiv preprint arXiv:2205.10839*, 2022. 1
- [45] Zhiyao Sun, Yu-Hui Wen, Tian Lv, Yanan Sun, Ziyang Zhang, Yaoyuan Wang, and Yong-Jin Liu. Continuously controllable facial expression editing in talking face videos. *arXiv preprint arXiv:2209.08289*, 2022. 2
- [46] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 716–731. Springer, 2020. 2
- [47] Guanzhong Tian, Yi Yuan, and Yong Liu. Audio2face: Generating speech/face animation from single audio with attention-based bidirectional lstm networks. In *2019 IEEE international conference on Multimedia & Expo Workshops (ICMEW)*, pages 366–371. IEEE, 2019. 2
- [48] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1415–1424, 2017. 3
- [49] 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. 2, 4, 12
- [50] Thurid Vogt, Elisabeth André, and Johannes Wagner. Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation. *Affect and Emotion in Human-Computer Interaction: From Theory to Applications*, pages 75–91, 2008. 2
- [51] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI*, pages 700–717. Springer, 2020. 6
- [52] Isabell Wohlgenannt, Alexander Simons, and Stefan Sieglitz. Virtual reality. *Business & Information Systems Engineering*, 62:455–461, 2020. 1

- [53] Tian Ye, Yunchen Zhang, Mingchao Jiang, Liang Chen, Yun Liu, Sixiang Chen, and Erkang Chen. Perceiving and modeling density for image dehazing. In *European Conference on Computer Vision*, pages 130–145. Springer, 2022. [1](#)
- [54] Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa. Speech emotion recognition using spectrogram & phoneme embedding. In *Interspeech*, volume 2018, pages 3688–3692, 2018. [3](#)
- [55] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019. [2](#)
- [56] Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3867–3876, 2021. [2](#)
- [57] Guangyan Zhang, Ying Qin, Wenjie Zhang, Jialun Wu, Mei Li, Yutao Gai, Feijun Jiang, and Tan Lee. iemotts: Toward robust cross-speaker emotion transfer and control for speech synthesis based on disentanglement between prosody and timbre. *arXiv preprint arXiv:2206.14866*, 2022. [3](#)
- [58] 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, 5, 13](#)
- [59] Kun Zhou, Berrak Sisman, and Haizhou Li. Vaw-gan for disentanglement and recombination of emotional elements in speech. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 415–422, 2021. [3](#)

Appendix

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.

A. 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.

A.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 [22]. 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.

A.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 $\mathbf{A}_{c1,e1}$ using $\mathbf{A}_{c1,e2}$ and $\mathbf{A}_{c2,e1}$ as inputs, then we need to interpolate them to have the same length as $\mathbf{A}_{c1,e1}$. After that, we decode the interpolated representation using 24 transformer[49] 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.

A.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[9] 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.

B. Blendshape to FLAME transform module

The Blendshape[26] to FLAME[27] 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. 6). 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:

$$V_{flame} = \sum_{i=1}^{52} \beta V_i \quad (7)$$

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

C. Baseline methods

We conducted a comparative analysis of EmoTalk with three state-of-the-art approaches, namely VOCA[8], MeshTalk[41], and FaceFormer[9]. 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.

D. 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[30] and HDTF[58] 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.

E. 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 [14] 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.

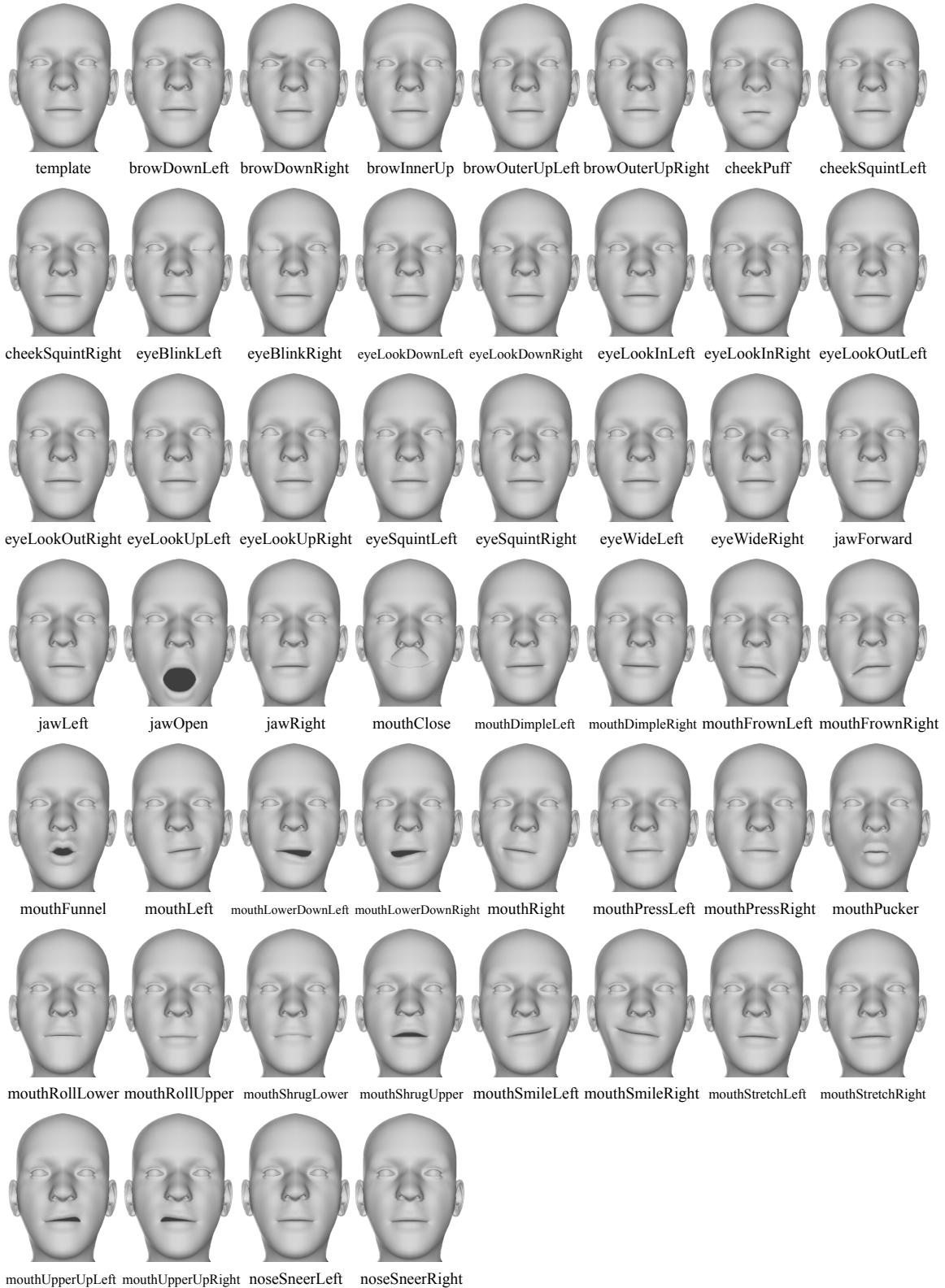


Figure 6. Semantically Meaningful FLAME Head Templates. We create 52 FLAME head templates that correspond to the blendshape coefficients, to achieve the transformation from the blendshape coefficients to the FLAME head model.