DIGITAL LIBRARY   Association for Computing Machinery   acm open>

RESEARCH-ARTICLE

PESTalk: Speech-Driven 3D Facial Animation with Personalized Emotional Styles

TIANSHUN HAN, Macau University of Science and Technology, Taipa, Macao

BENJIA ZHOU, Macau University of Science and Technology, Taipa, Macao

AJIAN LIU, Macau University of Science and Technology, Taipa, Macao

YANYAN LIANG, Macau University of Science and Technology, Taipa, Macao

DU ZHANG, Macau University of Science and Technology, Taipa, Macao

ZHEN LEI, Macau University of Science and Technology, Taipa, Macao

View all

Open Access Support provided by:

Macau University of Science and Technology

PESTalk: Speech-Driven 3D Facial Animation with Personalized Emotional Styles

Tianshun Han
Benjia Zhou
3230002542@student.must.edu.mo
zhou.benjia0228@gmail.com
Macau University of Science and
Technology
Macau, China

Ajian Liu
MAIS, CASIA
Beijing, China
Macau University of Science and
Technology
Macau, China
ajian.liu@ia.ac.cn

Yanyan Liang
Macau University of Science and
Technology
Macau, China
yyliang@must.edu.mo

Du Zhang
Macau University of Science and
Technology
Macau, China
duzhang@must.edu.mo

Zhen Lei
MAIS, CASIA
Beijing, China
Macau University of Science and
Technology
Macau, China
zhen.lei@ia.ac.cn

Jun Wan[†]
MAIS, CASIA
Beijing, China
Macau University of Science and
Technology
Macau, China
jun.wan@ia.ac.cn

Abstract

Speech-driven 3D facial animation aims to synthesize realistic emotional facial expressions that match the input speech. However, existing approaches are constrained by two key limitations: (1) These methods rely on pre-trained models (*e.g.*, Wav2Vec 2.0) as audio emotion feature extractors, which neglect critical frequency-domain characteristics, thereby emphasizing the challenge of discriminating between similar emotion categories. (2) They treat audio emotions as generic categorical states, ignoring individual differences in emotional expression, ultimately producing over-smoothed emotional representations that appear repetitive and stereotypical. To that end, we introduce **PESTalk**, a novel approach that generates 3D facial animations with **P**ersonalized **E**motional **S**tyles directly from speech inputs, thus significantly enhancing the realism of facial animations. Specifically, since acoustic frequency cues contain essential emotional information, we first propose a **D**ual-**S**tream **E**motion **E**xtractor (**DSEE**), which captures both time-domain variations and frequency-domain characteristics of audio signals to extract fine-grained affective features and subtle emotional nuances. Furthermore, we design an **E**motional **S**tyle **M**odeling **M**odule (**ESMM**) to achieve personalized emotional styles. This module first establishes a baseline representation for each subject based on voiceprint characteristics, then progressively refines it by continuously integrating emotional features. Ultimately, this process constructs a personalized emotional style representation for each subject in each emotion category, capturing their unique expression patterns. Finally, considering the scarcity of the 3D emotional talking face data, we employ an advanced facial capture model to extract pseudo facial blendshape coefficients from 2D emotional data, thereby constructing a large-scale 3D emotional talking face dataset with diverse emotions and personalized expressions (3D-EmoStyle). Extensive quantitative and qualitative evaluations show that PESTalk can generate realistic 3D facial animation and outperform state-of-the-art methods. The codes and dataset are available at: *https://github.com/tianshunhan/PESTalk*.

CCS Concepts

• **Computing methodologies → Computer vision**; **Rendering**.

Keywords

realistic 3D facial animation synthesis, personalized emotional styles, digital human

1 Introduction

Speech-driven 3D facial animation aims to generate realistic facial movements for 3D characters based on given speech input. Although there have been substantial advancements [9, 11, 16, 23, 24, 33] in lip synchronization, they neglect the overall facial expressions, resulting in only reliable lip movement. Consequently, the current researches [25, 31] is shifting towards creating comprehensive facial animations. Such animations are crucial for enhancing user experience, particularly in interactions with non-player characters in video games or emotionally responsive virtual chatbots [14].

However, recent methods [8, 12, 13, 25, 31] still suffer from two major limitations. First, most speech-driven 3D studies [8, 25, 31] rely on pre-trained speech models like Wav2Vec 2.0 [1] to extract
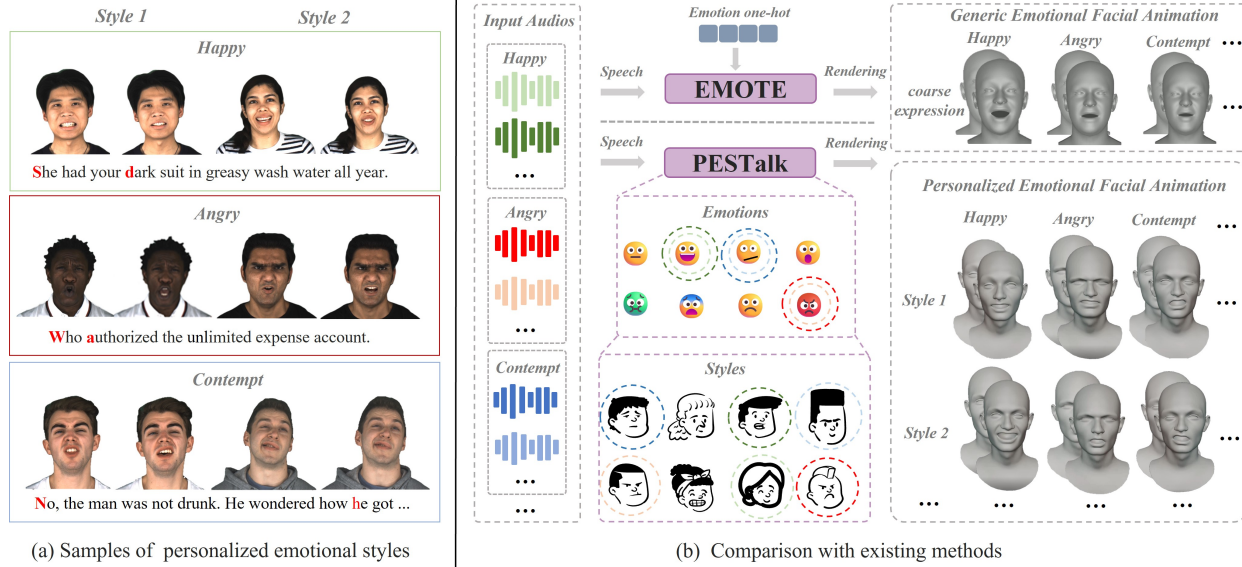
---

† Corresponding Author.

Style 1	Style 2		(b) Comparison with existing methods

(a) Samples of personalized emotional styles

Figure 1: (a) When expressing the same emotion with identical sentences, people show different facial expression patterns. This is influenced by habitual behaviors and cultural backgrounds, known as personalized emotional styles. (b) Prior work (*e.g.,* EMOTE [8]) treats emotion as one-hot embeddings, producing averaged expressions, while PESTalk dynamically selects optimal expressions by analyzing emotions and voiceprints, generating realistic 3D animations with personalized styles.

emotional information from input speech. Nevertheless, this approach relies on temporal convolutions to extract time-domain features, obscuring cross-frequency band correlations, thereby limiting its ability to discriminate between similar emotional categories [15, 21]. Moreover, some works [8, 13, 38] treat audio emotions as generic categorical states and represent them using one-hot vectors. Although this strategy offers an efficient way to regulate emotional expression in animation, it overlooks the personalized emotional expression, leading to over-smoothed facial animations. In reality, as illustrated in Fig. 1 (a), different speakers exhibit unique facial dynamics even when conveying the same emotion, highlighting that emotional expression is highly individualized [15, 21].

To address this challenge, we propose **PESTalk** (shown in Fig. 1 (b)), a novel speech-driven 3D facial animation model that extracts emotional information and personalized style cues from audio signals to generate highly realistic 3D facial expressions. Our framework comprises two essential modules: the Dual-Stream Emotion Extractor (DSEE), designed to enhance the capability of model in extracting fine-grained emotional features, and the Emotional Style Modeling Module (ESMM), which enables personalized emotional styles generation. Specifically, the DSEE consists of two independent audio encoders: a temporal stream encoder based on Temporal Convolutional Networks (TCN) to capture time-domain variations, and a frequency stream encoder utilizing mel-spectrogram analysis to extract frequency-domain characteristics. This complementary dual-stream mechanism enables a more comprehensive capture of subtle acoustic cues in emotional expression. Additionally, we develop the ESMM to enhance personalized expressiveness. The module first establishes baseline expression templates using speaker voiceprint characteristics, then dynamically constructs

a two-dimensional *<speaker × emotion>* emotional style library through continuously integrating emotional features of multiple emotional samples. During inference, the ESMM automatically retrieves the most matching personalized emotional style features from the input audio, thereby enabling the model with personalized expression capabilities. Finally, based on the biomechanical distinction that mouth movements (high-frequency for speech articulation) and upper-face motions (low-frequency for emotional expressions) require different processing, a partitioned style-guided fusion decoder is proposed to effectively handle this disparity.

To train the proposed network, emotional speeches with various personalized styles paired with corresponding 3D facial expressions are required. However, the absence of publicly suitable datasets poses a new challenge. To tackle this issue, we propose an emotional 3D talking face dataset, termed the 3D-EmoStyle dataset. To build this dataset, we use an advanced facial capture model [19] to extract "pseudo" 3D facial blendshape labels from 2D audio-visual datasets [37, 40]. Furthermore, to adapt the mesh-based algorithms, we employ deformation transfer [34] to create 52 FLAME head templates from the facial blendshape coefficients, allowing for efficient conversion between various facial animations.

In summary, our primary contributions are as follows:

- We propose a novel approach (PESTalk) to generate highly realistic 3D facial animation with personalized emotional styles, presenting state-of-the-art performance.
- We introduce a dual-stream emotion extractor (DSEE) to extract fine-grained emotional features and subtle nuances.
- We design an emotional style modeling module (ESMM) to enable personalized emotional styles generation.

- We collect a 3D talking face dataset (3D-EmoStyle) with diverse emotion categories and personalized expressions, including both facial blendshape and mesh vertex coefficients.

2 Related Work

2.1 Speech-Driven 3D Facial Animation

Previously, 2D facial animation methods [20, 22, 32, 41, 42] have garnered significant attention, primarily focusing on image-driven or speech-driven approaches to generate realistic videos of speaking individuals. However, these methods are not directly applicable to 3D character models, which are widely used in 3D games. As a result, speech-driven 3D facial animation [4, 23, 25, 31, 33] has recently emerged as a prominent research direction.

VOCA [6] pioneers the use of CNNs to map audio signals to 3D meshes, enabling the generation of facial animations with diverse speaking styles. MeshTalk [28] advances the field by focusing on full-face animations, successfully disentangling speech-correlated and speech-uncorrelated facial motions. FaceFormer [9] introduces a transformer-based model to generate continuous facial movements in an autoregressive manner, setting a new benchmark for temporal coherence. CodeTalker [39] further enriches the field by modeling the facial motion space using discrete primitives, achieving highly expressive animations. More recently, FaceDiffuser [33] becomes the first to apply probabilistic diffusion models to generate multiple facial motion options for a given speech, emphasizing lip-sync accuracy.

Despite these advancements, a critical limitation still persists. Current methods mainly concentrate on lip synchronization while overlooking the natural expression of emotions through facial animations. Capturing the subtle emotional nuances conveyed in speech is of great significance for enhancing the realism of 3D facial animations. In contrast, our work does not merely focus on lip synchronization but is also capable of achieving emotional expressions across the entire face.

2.2 Emotional 3D Facial Animation

The high coupling of content, emotion, and style in audio makes their decoupling a key challenge in realistic facial animation generation, drawing much research interest [8, 10, 25, 31]. EmoTalk [25] introduces an emotion disentanglement mechanism to extract emotional information from speech input, but its reliance on curated training data limits its effectiveness in zero-shot settings. DEITalk [31] addresses emotional saliency variations in long audio contexts by designing a dynamic emotional intensity module and a dynamic positional encoding strategy. EMOTE [8] employs a content-emotion exchange mechanism to supervise diverse emotions on the same audio while maintaining precise lip synchronization. Mimic [10] advances the field by learning disentangled representations of speaking style and content through two distinct latent spaces, enabling detailed modeling of 3D facial expressions.

Even with these developments, current methods often focus on facial emotional expressions while neglecting personalized emotional styles, resulting in less realistic animations. Additionally, most approaches are mesh-based, directly mapping speech signals to face mesh vertex coordinates, which can be challenging for artists and difficult to integrate into existing workflows. In this work, we

introduce a novel speech-driven 3D facial animation model that extracts emotional and style cues from audio to generate highly realistic expressions. By using facial blendshapes as the model output, we ensure seamless integration into industrial pipelines and enhance usability for professional animators.

3 Method

Overview. In this section, we detail the architectural components of our proposed framework, PESTalk (shown in Fig. 2), which generates highly realistic 3D facial animations with personalized emotional styles directly from speech inputs. The framework consists of two key modules: the Dual-Stream Emotion Extractor (DSEE, discussed in Section 3.1), which enhances the ability to extract fine-grained emotional features, and the Emotional Style Modeling Module (ESMM, covered in Section 3.2), which facilitates the generation of personalized emotional styles. Next, a style-guided fusion decoder (detailed in Section 3.3) transforms these features into facial animations. Additionally, we introduce a novel pairwise disentanglement mechanism (detailed in Section 3.4) to effectively separate content and emotion features of the input speech. The following subsections elaborate on the detailed design of each module.

Formulation. Let $B_{1:T} = (b_1, \ldots, b_T)$, $b_t \in \mathbb{R}^{52}$ be a T-length sequence of facial blendshape coefficients, which describes the ground truth of 3D face movements. Let $A_{1:T} = (a_1, \ldots, a_T)$ be a sequence of speech snippets, and each $a_t \in \mathbb{R}^D$ has D samples to align with the corresponding (visual) frame b_t. The proposed model takes speech $A_{1:T}$ as input, predicting facial blendshape coefficients $\hat{B}_{1:T} = \left(\hat{b}_1, \ldots, \hat{b}_T \right)$. Formally,
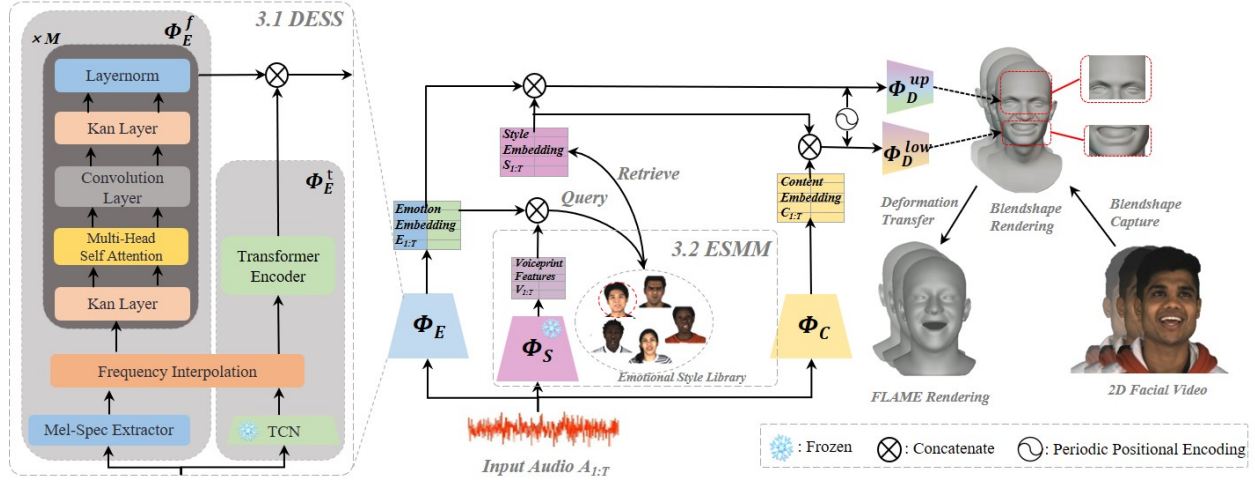
$$\hat{b}_t = \text{PESTalk}_\theta \left(a_t \right), \tag{1}$$

where θ indicates the model parameters, t is the current time-step in the sequence and $\hat{b}_t \in \hat{B}_{1:T}$.

3.1 Dual-Stream Emotion Extractor

Prior works [8, 14, 25] primarily utilize pre-trained speech models as emotion encoders, which fail to adequately capture frequency-domain characteristics and often confuse similar emotional categories [15, 21]. To overcome this limitation, we propose a novel dual-stream emotion extractor Φ_E (see Fig. 2 left) that synergistically models both time-domain and frequency-domain emotional cues. The temporal stream Φ_E^t is a pre-trained audio model fine-tuned on emotion [1]. It is composed of several frozen TCN layers and multi-layer transformer blocks to extract robust time-domain features. The frequency stream Φ_E^f incorporates a KAN-enhanced Conformer architecture that replaces traditional MLPs with KAN layers [17] which employs spline-parameterized activation functions, enabling more precise modeling of complex emotional patterns in the frequency domain. This hybrid architecture simultaneously captures both long-term dependencies and short-term variations.

Specifically, for an input audio segment $A_{1:T} = (a_1, \ldots, a_T)$, we first extract time-domain features through TCN layers and compute mel-spectrogram features. After interpolation transformations, we process them through parallel pathways: the frequency features undergo KAN-Conformer blocks to extract $E_{f_{1:T}}$, while the temporal features pass through a transformer encoder to obtain $E_{t_{1:T}}$. These

Figure 2: Overview of PESTalk. Given a speech input $A_{1:T}$, PESTalk extracts emotion features $E_{1:T}$ and voiceprint information $V_{1:T}$. Then, it uses these two types of information to query the emotional style library and match the closest personalized emotional style $S_{1:T}$. After that, PESTalk extracts content features $C_{1:T}$ from the speech and integrates these features. Subsequently, two distinct sets of decoders, Φ_D^{up} and Φ_D^{low}, combine these integrated features to generate facial blendshape coefficients for the upper and lower face, respectively. Additionally, these coefficients also can be used to animate a FLAME model.

are concatenated $[E_{t_{1:T}} \parallel E_{f_{1:T}}]$ and linearly projected via Φ_E^{proj} to form a joint representation $E_{1:T}$:

$$
\begin{aligned}
E_{t_{1:T}} &= \Phi_E^t(A_{1:T}), \quad E_{f_{1:T}} = \Phi_E^f(A_{1:T}), \\
E_{1:T} &= \Phi_E^{\text{proj}}([E_{t_{1:T}} \parallel E_{f_{1:T}}]), \quad E_t \in \mathbb{R}^{256},
\end{aligned}
\tag{2}
$$

where $E_{1:T}$ integrates both temporal dynamics and spectral characteristics for comprehensive emotion representation, and $[\cdot \parallel \cdot]$ represents the concatenation operation.

3.2 Emotional Style Modeling Module

Emotional Styel Library Construction. Voiceprint information demonstrates a clear correlation with individuals, as it is shaped by unique physiological structures and neuromuscular control patterns, resulting in high individual specificity [27]. Based on this, for each subject, we use its voiceprint information as the base style representation. Specifically, there are K subjects in the training set, and each subject has N audio clips. We utilize a pre-trained speaker recognition model [2, 5] to extract the voiceprint features of each audio clip. We average the voiceprint features of all audio clips for the same subject, obtaining the base style representation, denoted as $\mathbf{R}_i$, where $i = 1, 2, \cdots, K$. Formally, we define:

$$
\mathbf{R}_i = \frac{1}{N} \sum_{j=1}^N \mathbf{V}_{ij}, \quad \mathbf{R}_i \in \mathbb{R}^{512},
$$

Here, $\mathbf{V}_{ij}$ represents the voiceprint feature of the j-th audio clip of the i-th subject.

However, the above base style representation is difficult to establish a connection with the emotional information of the audio. During the training process, each subject has C emotion categories, and each emotion category contains X audio clips. For each subject, we compute the average of the emotional features $E_{1:T}$ across all

audio clips belonging to the same emotion category. Subsequently, the averaged features is then concatenated with the base style representation $\mathbf{R}_i$ to generate a personalized emotional style representation, which is denoted as $\mathbf{P}_{ic}$, with $c = 1, 2, \cdots, C$. The formula is expressed as:

$$
\mathbf{P}_{ic} = [\mathbf{R}_i \parallel \frac{1}{X} \sum_{k=1}^X \mathbf{E}_{ick}], \quad \mathbf{P}_{ic} \in \mathbb{R}^{768},
$$

Here, $\mathbf{E}_{ick}$ represents the emotional feature of the k-th audio clip of the i-th subject under the c-th emotion category. In this way, we dynamically construct a two-dimensional *<speaker × emotion>* emotional style library $\mathcal{L}$.

Personalized Emotional Style Retrieval. Individuals with similar voiceprint characteristics exhibit convergent expressive styles, since voiceprint parameters encode both anatomical similarities and consistency in vocal habits [30]. During the retrieval process, for an input speech $A_{1:T} = (a_1, \ldots, a_T)$, it is passed through Φ_E and Φ_S respectively to obtain the emotional feature E_{ic} and voiceprint feature R_i, which are then concatenated. The concatenated result is used to query in the pre-built emotional style library $\mathcal{L}$ to find the closest personalized emotional style representation. Specifically, the optimal style embedding is obtained via maximum similarity retrieval, which is expressed as:

$$
S_i = \arg\min_{S_k \in \mathcal{L}} D_{\cos}([E_{ic} \parallel R_i], S_k), \quad S_i \in \mathbb{R}^{768}, \tag{3}
$$

$$
D_{\cos}(u, v) = 1 - \frac{u^\top v}{\|u\|\|v\|}, \quad \text{where } u = [E_{ic} \parallel R_i], \ v = S_k, \tag{4}
$$

Here, S_k represents an arbitrary emotional style feature in the emotional style library $\mathcal{L}$. The derived S_i denotes the closest personalized emotional style to the input speech. Before being fed into

subsequent processing stages, the retrieved features S_i are passed through a linear projection layer to obtain a lower-dimensional representation.

3.3 Emotion-content Disentanglement

Content Extractor. We employ the pretained Wav2Vec 2.0 model [1] as our content encoder Φ_C, building upon recent advances in speech-driven animation [9, 10, 25, 31]. The encoder processes raw audio $A_{1:T}$ through temporal convolutional layers Φ_C^{conv} followed by a transformer encoder Φ_C^{trans} with multi-head self-attention [36], and finally a linear projection layer Φ_C^{proj} for dimensional adjustment. This pipeline generates frame-level content features:

$$C_{1:T} = \Phi_C(a_1, \ldots, a_T), \quad c_t \in \mathbb{R}^{256}, \tag{5}$$

We freeze Φ_C^{conv} while fine-tuning both Φ_C^{trans} and Φ_C^{proj}, maintaining stable acoustic feature extraction while adapting to our facial animation domain through learnable contextual modeling.

Pairwise Disentanglement Mechanism. To enhance the model's ability to disentangle content features and emotional features, we proposes a new disentanglement mechanism (illustrated in Figure. 3). Compared with the strategy of forced cross-reconstruction adopted in previous work [8, 25], this mechanism does not require strict data pairing conditions and can achieve feature separation with a more concise architecture.

Specifically, we input two audio clips that have the same content but differ in emotion, with one being neutral and the other emotional, denoted as $\{(A^{\text{neutral}}, A^{\text{emotion}})\}$. The content extractor and emotion extractor then project these paired audios into content and emotion latent spaces, respectively, pulling semantically identical features closer while pushing emotionally distinct ones apart. Formally:
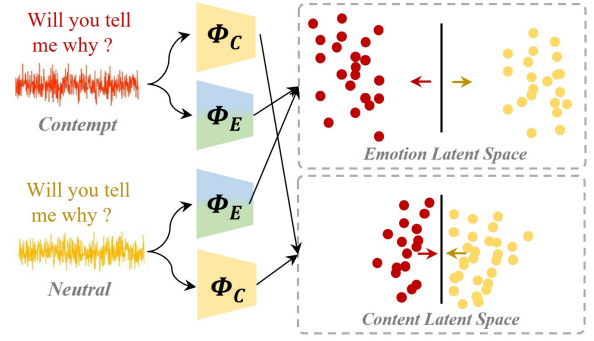
$$
\begin{aligned}
z_i^c &= \Phi_C(A_i^{\text{neutral}}), \quad z_i^e = \Phi_E(A_i^{\text{neutral}}), \\
z_j^c &= \Phi_C(A_i^{\text{emotion}}), \quad z_j^e = \Phi_E(A_i^{\text{emotion}}), \\
\mathcal{L}_{\text{dis}} &= \sum_{i,j} \|z_i^c - z_j^c\|_\psi + \sum_{i,j} \max(0, \delta - \|z_i^e - z_j^e\|),
\end{aligned}
\tag{6}
$$

where ψ-norm enforces content compactness and δ denotes the learnable margin threshold. Through this gentle decoupling method, enabling the model to learn discriminative features that distinguish between emotional and neutral expressions.

3.4 Style-Guided Animation Generation

The human face exhibits distinct motion patterns between upper and lower regions [14, 31]. The lower face displays high-frequency movements tightly coupled with speech articulation, while the upper face shows sustained, low-frequency motions reflecting emotional states. To model this biomechanical dichotomy, we propose a partitioned style-guided fusion decoder: a lower-face decoder Φ_D^{low} that integrates phonetic content $C_{1:T}$, and personalized style $S_{1:T}$ for precise speech-synchronized animation; and an upper-face decoder Φ_D^{up} that processes emotional $E_{1:T}$ and personalized style $S_{1:T}$ to generate emotion-driven facail motions. This specialized design enables accurate modeling of both speech-related articulations and sustained emotional expressions.

Figure 3: Pairwise Disentanglement Mechanism. The content extractor Φ_C and emotion extractor Φ_E project paired audios (same semantics, different emotions) into content and emotion latent spaces, respectively. The mechanism pulls close semantically identical features while pushing apart emotionally distinct ones.

Both decoders employ an identical processing pipeline. First, periodic positional encoding [9] injects temporal regularity crucial for speech-synchronized lip movements. Next, a biased multi-head self-attention layer (inspired by ALiBi [26]) computes context-aware feature representations, where the attention bias mechanism prioritizes local temporal dependencies through its decaying mask pattern. The resultant features undergo transformation via feed-forward networks before final generation to facial blendshape. The architecture outputs 32 lower-face blendshapes and 20 upper-face blendshapes, which concatenate into 52-dimensional facial blendshape coefficients.

3.5 Objective function

We utilize a comprehensive loss function that encompasses four key components: position loss, motion loss, classification loss, and disentanglement loss. The complete function is defined as follows:

$$L = \lambda_1 L_{\text{pos}} + \lambda_2 L_{\text{mot}} + \lambda_3 L_{\text{cls}} + \lambda_4 L_{\text{dis}}, \tag{7}$$

where $\lambda_1 = 1$, $\lambda_2 = 0.5$, $\lambda_3 = 0.1$, and $\lambda_4 = 0.01$ in all of our experiments. A comprehensive explanation of each of these components is provided in the following section. Please note that in the follow equations, N denotes the number of samples in the training set, T denotes the length of each input sequence.

Position Loss. The position loss measures the difference between the predicted facial blendshape coefficients and the corresponding ground truth facial blendshape coefficients. Specifically, we use per-frame mean squared error (MSE) as the position loss:

$$L_{\text{pos}} = \frac{1}{NT} \sum_i^N \sum_{t=1}^T \left\| \left(\hat{B}_{1:T} - B_{1:T} \right) \right\|^2. \tag{8}$$

Motion Loss. To address jittery output frames, we introduce a motion loss that ensures temporal stability. This loss considers the smoothness between predicted and ground truth frames:

$$L_{mot} = \frac{1}{NT} \sum_i^N \sum_{t=1}^T \left\| \left(\hat{B_{2:T-1}} - \hat{B_{1:T}} \right) - \left(B_{2:T-1} - B_{1:T} \right) \right\|^2. \tag{9}$$

Table 1: Quantitative evaluation results on blendshape-based datasets. Best performance highlighted in bold and lower values indicate better performance for metrics marked with ↓.

Method	CREMA-D				RAVDESS				3D-EmoStyle			
	LBE↓	PBE↓	MBE↓	BA↑	LBE↓	PBE↓	MBE↓	BA↑	LBE↓	PBE↓	MBE↓	BA↑
EmoTalk [25]	0.208	0.287	0.374	0.713	0.166	0.242	0.371	0.752	0.199	0.259	0.313	0.642
FaceDiffuser [33]	**0.168**	**0.254**	0.753	0.742	0.152	0.266	0.416	0.721	0.178	0.218	0.458	0.627
Ours	0.179	0.262	**0.336**	**0.758**	**0.145**	**0.227**	**0.362**	**0.771**	**0.130**	**0.190**	**0.277**	**0.739**

Table 2: Quantitative evaluation results on vertex-based datasets.

Method	CREMA-D			RAVDESS			3D-EmoStyle		
	LVE↓$_{\times 10^{-3}}$	EVE↓$_{\times 10^{-4}}$	FDD↓$_{\times 10^{-4}}$	LVE↓$_{\times 10^{-3}}$	EVE↓$_{\times 10^{-4}}$	FDD↓$_{\times 10^{-4}}$	LVE↓$_{\times 10^{-3}}$	EVE↓$_{\times 10^{-4}}$	FDD↓$_{\times 10^{-4}}$
EmoTalk [25]	4.829	8.605	4.633	6.823	8.081	4.469	3.544	9.848	4.565
FaceDiffuser [33]	**3.871**	9.562	4.806	3.423	9.351	4.865	3.498	14.486	8.196
DEEPTalk [14]	4.127	8.138	3.733	3.833	7.428	3.239	3.516	8.462	3.853
Ours	3.931	**7.638**	**3.208**	**3.236**	**6.593**	**3.153**	**3.183**	**6.517**	**2.771**

Classification Loss. We introduce a classification loss to supervise the output of the emotion extractor Φ_E. The classification loss is defined as follows:

$$L_{cls} = -\frac{1}{N} \sum_i^N \sum_{c=1}^M \left(\boldsymbol{y}_{ic} * \log \boldsymbol{p}_{ic} \right), \qquad (10)$$

where M denotes the number of classifications, $\boldsymbol{y}_{ic}$ is the observation function that determines whether the sample i carries the emotion label c, and $\boldsymbol{p}_{ic}$ denotes the predicted probability that sample i belongs to class c.
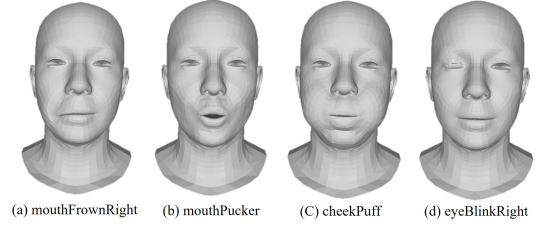
Disentanglement Loss. Our training framework processes content-parallel audio pairs $(A_{1:T}, \hat{A}_{1:T})$ with different emotional states—one neutral, one expressive. The model aims to maintain content via feature alignment and separate emotions via feature separation. The disentanglement mechanism is defined as:

$$
\begin{aligned}
E &= \text{Pool}(\Phi_E(A_{1:T})), \quad \hat{E} = \text{Pool}(\Phi_E(\hat{A}_{1:T})), \\
C &= \text{Pool}(\Phi_C(A_{1:T})), \quad \hat{C} = \text{Pool}(\Phi_C(\hat{A}_{1:T})), \\
\mathcal{L}_{\text{dis}} &= 1 - \frac{1}{N} \sum_{i=1}^N \left[\cos(E_i, \hat{E}_i) - \cos(C_i, \hat{C}_i) \right],
\end{aligned} \qquad (11)
$$

where $\text{Pool}(\cdot)$ aggregates frame-level features via average pooling, and $\cos(\cdot, \cdot)$ measures cosine similarity.

4 Experiments

Datasets. Due to the challenges in manually collecting high-quality emotional facial blendshape data from actors, we employ an established facial capture model [19] to extract pseudo-blendshape coefficients from large-scale emotional 2D video datasets (MEAD [37] and HDTF [40]), thereby constructing the 3D-EmoStyle dataset. This dataset comprises 180+ subjects exhibiting 8 prototypical emotional expressions. To facilitate comparative analysis with mesh-based approaches, we further convert the blendshape coefficients to mesh vertex data using the deformation transfer [34], generating

(a) mouthFrownRight (b) mouthPucker (C) cheekPuff (d) eyeBlinkRight
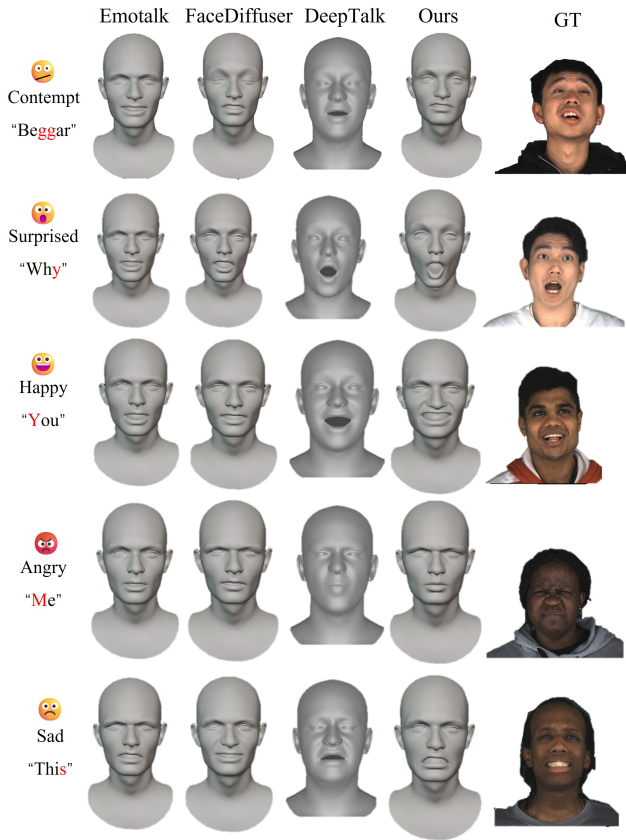
Figure 4: Samples of reference meshes from facial blendshape coefficients to FLAME mesh.

mesh-based labels (shown in Figure 4). However, since MEAD's overlapping utterances between training and test sets compromise fair evaluation, we additionally incorporated CREMA-D [3] and RAVDESS [18] datasets for assessment.

Baseline Implementations. Since our model aims to achieve personalized 3D facial emotion generation, we focus on comparing current emotion-based speech-driven methods for a fair comparison. Specifically, we conduct comprehensive comparisons between PESTalk and three state-of-the-art open-source approaches, EmoTalk [25], FaceDiffuser [33], and DEEPTalk [14]. For DEEPTalk, we follow the methodology of the original paper by using EMO-CAv2 [7] mdeol to extract FLAME parameters for model training. All models are trained on the training set of the 3D-EmoStyle dataset and subsequently evaluated on its test set along with two additional benchmark datasets.

4.1 Quantitative Evaluation

Evaluation Metrics. We employ a comprehensive set of evaluation using both blendshape-based and vertex-based metrics. For blendshape analysis, we adopt LBE (Lip Blendshape Error) [33] to measure lip-sync accuracy, while introducing PBE (Pronunciation Blendshape Error) to specifically evaluate pronunciation-related facial movements (*e.g.* mouth and jaw). Additionally, MBE (Mean

Figure 5: Visual comparison of facial movements generated by different methods on the 3D-EmoStyle test set. The results demonstrate samples of distinct emotional expressions. Compared with other approaches, our method produces more emotionally expressive and realistic facial animations.

Figure 6: Render results in Unreal Engine. The facial blend-shapes generated by PESTalk are used to drive a digital human, which is then rendered in Unreal Engine.

Blendshape Error) [33] is used to quantify the overall deviation across all blendshape parameters by computing the maximal L2 error per frame and averaging across the sequence. To assess temporal synchronization, we compute BA (Beat Alignment) [35] between predicted and ground truth outputs. For vertex-based evaluation, we utilize three metrics: LVE (Lip Vertex Error) [9, 29] calculates the maximum L2 error among lip vertices to measure lip synchronization accuracy; EVE (Emotional Vertex Error) [25] extends this evaluation to around the eyes and forehead regions for comprehensive expression assessment; and FDD (Upper Face Dynamics Deviation) [39] compares the standard deviation of upper face vertex motions over time.

Evaluation on Lip Synchronization. As shown in Table 1, the PESTalk achieved remarkable performance across all tested datasets in three key metrics: LBE, PBE, and MBE. This robust performance demonstrates the model's strong capability of precise lip synchronization and generalization. Notably, PESTalk also attained the best scores in the BA metric. The BA metric evaluates synchronization by comparing audio beats with corresponding facial movement

patterns derived from blendshape coefficients. Since lip motions occur at high frequencies, this can reflect the lip synchronization. These results further validate the exceptional lip-sync precision of PESTalk. Furthermore, while previous methods [9, 29, 39] often employ the LVE metric for lip-sync evaluation, our model directly outputs facial blendshape coefficients that cannot be directly compared. To address this, as presented in Table 2, we converted the output blendshape coefficients into mesh vertex coordinates for comparison with ground truth data. Experimental results show that our model achieved the lowest LVE error across all datasets, providing additional compelling evidence for its accurate lip synchronization capability.
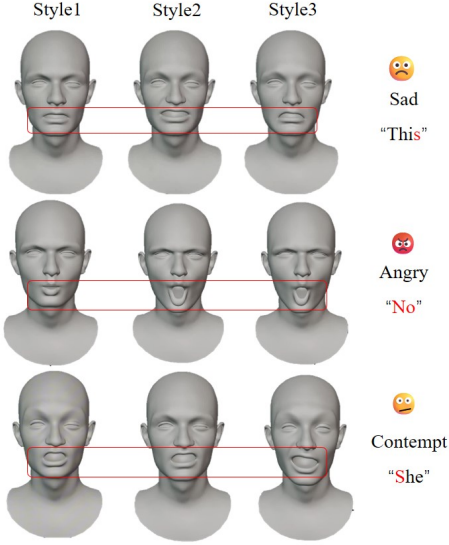
Evaluation on Emotional Expression. In evaluating emotional expression, we employed three key metrics: FDD to analyze discrepancies in upper facial movements compared to ground truth data, EVE to assess subtle expression variations in around the eyes and forehead regions (such as furrowed brows in anger expressions), and MBE to evaluate overall facial deviations. Experimental results demonstrate that PESTalk significantly outperforms all other methods across all datasets in terms of FDD, EVE, and MBE metrics (see Tables 1 and 2), conclusively validating its superior capability in emotional expression reproduction and its precision in capturing natural facial dynamics and nuanced expression details.

4.2 Qualitative Evaluation

Visual Comparison. In the comparative experiments shown in Figure 5, we evaluated PESTalk against several state-of-the-art methods using proposed 3D-EmoStyle test set. The experimental results demonstrate that while most existing methods can generate relatively natural lip movements, they still exhibit significant shortcomings in detailed performance. Specifically, EmoTalk and FaceDiffuser show inadequate performance in speech synchronization accuracy. Taking the /aɪ/ phoneme in the word "why" as an example, these two methods fail to accurately reproduce the required articulatory features (the natural jaw drop and full mouth opening). In contrast, our model is capable of generating precisely lip-synced animations.

In terms of emotional expression, as shown in Figure 5, it can be observed that except for DEEPTalk, other methods often produce incorrect or expressionless results that deviate from ground truth references. While DEEPTalk demonstrates improved emotional expressiveness, it still exhibits significant limitations. The method performs poorly in recognizing complex emotions such as "contempt," resulting in animations that substantially differ from

Figure 7: The personalized emotional expression results of PESTalk. We evaluated the performance across samples of three distinct emotional states, where different individuals speaking the same sentence.

the ground truth. In contrast, our approach consistently generates emotionally expressive and highly realistic facial animations across all emotional categories.

Additionally, PESTalk outputs facial blendshapes, enabling seamless integration into industry pipelines and enhancing usability for professional animators [25, 31]. For instance, as illustrated in Figure 6, we use MetaHuman4 to create our digital avatar and Unreal Engine 5 to render its animations. The animations are driven by the facial blendshape coefficients produced by our model. This approach allows our model to drive any digital character according to the rendering engine's rules, showcasing its strong practicality.

Personalized Emotional Expression. We constructed test groups comprising audio samples with identical semantic content and emotion category but from different speakers. As shown in Figure 7, the experimental results demonstrate that despite sharing identical emotion labels and textual content, PESTalk can accurately match personalized emotional expression styles based on speaker characteristics, generating highly realistic and individualized facial animations. This personalization is particularly in contempt expressions where noticeable differences existed in both the direction and degree of lip asymmetry among subjects.

4.3 User Studies

To conduct a more comprehensive evaluation of our approach, we designed a user study with the following experimental protocol. Specifically, we randomly selected 5 samples from each emotion category of every subject in the 3D-EmoStyle test set. Then, we invited 20 participants to compare the results of all models with the ground truth. The participants evaluated the results in two dimensions (lip-sync accuracy and personalized emotional expressiveness) using a Mean Opinion Score (MOS) ranging from 1 to 5,

Table 3: User Study Results. Our method demonstrates superior performance compared to SOTA about personalized emotional expression (PEE) and lip synchronization (Lip Sync).

Method	PEE↑	Lip Sync↑
EmoTalk [25]	3.878	3.598
FaceDiffuser [33]	3.199	3.862
DEEPTalk [14]	4.130	3.981
Ours	**4.349**	**4.276**

Table 4: Ablation results on the 3D-EmoStyle test set.

Method	LBE↓	MBE↓	EVE↓$_{\times 10^{-4}}$	FDD↓$_{\times 10^{-4}}$
w/o Φ_E^f	0.143	0.302	8.763	3.562
w/o ESMM	0.149	0.317	9.202	4.103
w/o $\mathcal{L}_{dis}$	0.169	0.375	9.943	4.279
Full (Ours)	**0.130**	**0.277**	**6.517**	**2.771**

where a higher score indicates better performance. As presented in Table 3, the results clearly show that our method substantially outperforms existing approaches in terms of both lip-synchronization precision and personalized emotional expression.

4.4 Ablation Studies

We conducted comprehensive ablation studies on the 3D-EmoStyle test set to systematically evaluate the contributions of individual components in PESTalk. As demonstrated in Table 4, removing the frequency stream of DESS (w/o Φ_E^f) significantly impairs the emotional expressiveness, evidenced by concurrent degradation in EVE and FDD. Similarly, the absence of ESMM compromises personalized emotional expression capabilities, which results in the output tending to be consistent for the same emotion category. Most notably, eliminating the disentanglement loss (w/o $\mathcal{L}_{dis}$) results in substantial performance drops across all metrics, as the model fails to effectively disentangle speech content from emotional features, thereby adversely affecting both lip-sync accuracy and emotional expression quality.

5 Conclusion

This paper introduces PESTalk, a novel speech-driven 3D facial animation framework that generates personalized emotional expressions while maintaining accurate lip synchronization. Unlike conventional approaches, PESTalk achieves this through two key components: (1) a Dual-Stream Emotion Extractor (DSEE) that captures both time-domain and frequency-domain emotional cues for fine-grained feature extraction, and (2) an Emotional Style Modeling Module (ESMM) that enables personalized emotional styles generation. To address data scarcity, we construct the 3D-EmoStyle, a large-scale 3D emotional talking face dataset with diverse personalized expressions. Extensive experiments on various datasets demonstrate the superiority of our model over existing methods across seven metrics, with notable improvements in emotional authenticity and personalized expression.

Acknowledgments

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.

[2] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 7124–7128.

[3] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing* 5, 4 (2014), 377–390.

[4] Peng Chen, Xiaobao Wei, Ming Lu, Yitong Zhu, Naiming Yao, Xingyu Xiao, and Hui Chen. 2023. DiffusionTalker: Personalization and Acceleration for Speech-Driven 3D Face Diffuser. *arXiv preprint arXiv:2311.16565* (2023).

[5] Juan M. Coria, Hervé Bredin, Sahar Ghannay, and Sophie Rosset. 2020. A Comparison of Metric Learning Loss Functions for End-To-End Speaker Verification. In *Statistical Language and Speech Processing*, Luis Espinosa-Anke, Carlos Martín-Vide, and Irena Spasić (Eds.). Springer International Publishing, 137–148.

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

[7] Radek Daněček, Michael J Black, and Timo Bolkart. 2022. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 20311–20322.

[8] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. 2023. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*. 1–13.

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

[10] Hui Fu, Zeqing Wang, Ke Gong, Keze Wang, Tianshui Chen, Haojie Li, Haifeng Zeng, and Wenxiong Kang. 2024. Mimic: Speaking Style Disentanglement for Speech-Driven 3D Facial Animation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 1770–1777.

[11] Tianshun Han, Shengnan Gui, Yiqing Huang, Baihui Li, Lijian Liu, Benjia Zhou, Ning Jiang, Quan Lu, Ruicong Zhi, Yanyan Liang, et al. 2024. PMMTalk : Speech-Driven 3D Facial Animation from Complementary Pseudo Multi-modal Features. *IEEE Transactions on Multimedia* (2024).

[12] Kazi Injamamul Haque and Zerrin Yumak. 2023. FaceXHuBERT: Text-less Speech-driven E (X) pressive 3D Facial Animation Synthesis Using Self-Supervised Speech Representation Learning. *arXiv preprint arXiv:2303.05416* (2023).

[13] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–12.

[14] Jisoo Kim, Jungbin Cho, Joonho Park, Soonmin Hwang, Da Eun Kim, Geon Kim, and Youngjae Yu. 2024. DEEPTalk: Dynamic Emotion Embedding for Probabilistic Speech-Driven 3D Face Animation. *arXiv preprint arXiv:2408.06010* (2024).

[15] Yante Li, Jinsheng Wei, Yang Liu, Janne Kauttonen, and Guoying Zhao. 2022. Deep learning for micro-expression recognition: A survey. *IEEE Transactions on Affective Computing* 13, 4 (2022), 2028–2046.

[16] Yihong Lin, Zhaoxin Fan, Xianjia Wu, Lingyu Xiong, Liang Peng, Xiandong Li, Wenxiong Kang, Songju Lei, and Huang Xu. 2024. Glditalker: Speech-driven 3d facial animation with graph latent diffusion transformer. *arXiv preprint arXiv:2408.01826* (2024).

[17] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. 2024. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756* (2024).

[18] Steven R Livingstone and Frank A Russo. 2018. 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 (2018), e0196391.

[19] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).

[20] Haoyu Ma, Tong Zhang, Shanlin Sun, Xiangyi Yan, Kun Han, and Xiaohui Xie. 2024. CVTHead: One-shot controllable head avatar with vertex-feature transformer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 6131–6141.

[21] David Matsumoto. 2013. Culture and emotional expression. *Understanding Culture* (2013), 271–287.

[22] Fatemeh Nazarieh, Zhenhua Feng, Diptesh Kanojia, Muhammad Awais, and Josef Kittler. 2024. PortraitTalk: Towards Customizable One-Shot Audio-to-Talking Face Generation. *arXiv preprint arXiv:2412.07754* (2024).

[23] Inkyu Park and Jaewoong Cho. 2023. SAiD: Speech-Driven Blendshape Facial Animation with Diffusion. *arXiv preprint arXiv:2401.08655* (2023).

[24] Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. 2023. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5292–5301.

[25] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. 2023. EmoTalk: Speech-driven emotional disentanglement for 3D face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20687–20697.

[26] Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409* (2021).

[27] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. 2000. Speaker verification using adapted Gaussian mixture models. *Digital signal processing* 10, 1-3 (2000), 19–41.

[28] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. 2021. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1173–1182.

[29] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. 2021. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1173–1182.

[30] Stallone B Sabatier, Morgan R Trester, and Jeremy M Dawson. 2019. Measurement of the impact of identical twin voices on automatic speaker recognition. *Measurement* 134 (2019), 385–389.

[31] Kang Shen, Haifeng Xia, Guangxing Geng, Guangyue Geng, Siyu Xia, and Zhengming Ding. 2024. DEITalk: Speech-Driven 3D Facial Animation with Dynamic Emotional Intensity Modeling. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 10506–10514.

[32] Zhicheng Sheng, Liqiang Nie, Meng Liu, Yinwei Wei, and Zan Gao. 2023. Towards Fine-Grained Talking Face Generation. *IEEE Transactions on Image Processing* (2023).

[33] Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. 2023. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*. 1–11.

[34] Robert W Sumner and Jovan Popović. 2004. Deformation transfer for triangle meshes. *ACM Transactions on graphics (TOG)* 23, 3 (2004), 399–405.

[35] Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong-jin Liu. 2024. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–9.

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

[37] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*. Springer, 700–717.

[38] Qianyun Wang, Zhenfeng Fan, and Shihong Xia. 2021. 3d-talkemo: Learning to synthesize 3d emotional talking head. *arXiv preprint arXiv:2104.12051* (2021).

[39] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. 2023. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 12780–12790.

[40] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 2021. 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*. IEEE, 3661–3670.

[41] Longtao Zheng, Yifan Zhang, Hanzhong Guo, Jiachun Pan, Zhenxiong Tan, Jiahao Lu, Chuanxin Tang, Bo An, and Shuicheng Yan. 2024. MEMO: Memory-Guided Diffusion for Expressive Talking Video Generation. *arXiv preprint arXiv:2412.04448* (2024).

[42] Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, and Guanbin Li. 2023. Identity-preserving talking face generation with landmark and appearance priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 9729–9738.