Speech-Driven 3D Face Animation with Composite and Regional Facial Movements

Haozhe Wu
wuhz19@mails.tsinghua.edu.cn
Department of Computer Science and Technology, Tsinghua University
Beijing 100084, China

Songtao Zhou
zhoust19@mails.tsinghua.edu.cn
Department of Computer Science and Technology, Tsinghua University
Beijing 100084, China

Jia Jia[*]
jjia@tsinghua.edu.cn
Department of Computer Science and Technology, Tsinghua University
Beijing National Research Center for Information Science and Technology
Beijing 100084, China

Junliang Xing
jlxing@tsinghua.edu.cn
Department of Computer Science and Technology, Tsinghua University
Beijing 100084, China

Qi Wen
2838158073@qq.com
ByteDance
Hangzhou, China

Xiang Wen
wenxiang@zju.edu.cn
ByteDance
Hangzhou, China

ABSTRACT

Speech-driven 3D face animation poses significant challenges due to the intricacy and variability inherent in human facial movements. This paper emphasizes the importance of considering both the composite and regional natures of facial movements in speech-driven 3D face animation. The composite nature pertains to how speech-independent factors globally modulate speech-driven facial movements along the temporal dimension. Meanwhile, the regional nature alludes to the notion that facial movements are not globally correlated but are actuated by local musculature along the spatial dimension. It is thus indispensable to incorporate both natures for engendering vivid animation. To address the composite nature, we introduce an adaptive modulation module that employs arbitrary facial movements to dynamically adjust speech-driven facial movements across frames on a global scale. To accommodate the regional nature, our approach ensures that each constituent of the facial features for every frame focuses on the local spatial movements of 3D faces. Moreover, we present a non-autoregressive backbone for translating audio to 3D facial movements, which maintains high-frequency nuances of facial movements and facilitates efficient inference. Comprehensive experiments and user studies demonstrate that our method surpasses contemporary state-of-the-art approaches both qualitatively and quantitatively.

CCS CONCEPTS

• **Computing methodologies → Shape representations**; **Animation**.

KEYWORDS

Speech-Driven 3D Face Animation, Regional, Composite, Non-autoregressive

ACM Reference Format:
Haozhe Wu, Songtao Zhou, Jia Jia, Junliang Xing, Qi Wen, and Xiang Wen. 2023. Speech-Driven 3D Face Animation with Composite and Regional Facial Movements . In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3611775

1 INTRODUCTION

Speech-driven 3D face animation is a crucial technology in the field of digital avatar synthesis, which has wide-ranging applications in VR/AR, games, and film-making. However, the intricacy and variability of human facial movements pose significant challenges for this task. Human facial movements have two essential natures: composite and regional. The composite nature refers to how speech-independent factors, such as talking styles and expressions, globally modulate speech-driven facial movements along the temporal dimension. For example, different talking styles can affect the amplitude of mouth opening while speaking the same word. The regional nature arises because the movement of different facial parts is not globally connected along the spatial dimension but rather determined by the action of local muscles. For instance, the movements of the eyebrows are usually uncorrelated with those of the jaw. Understanding and modeling composite and regional natures are crucial for realistic and vivid facial animation.

Several previous studies [10, 12, 32, 40] have attempted to incorporate the composite nature into speech-driven 3D face animation models by fusing speech-independent labels such as emotion, identity, and style. However, these fused labels are often coarse-grained, which limits their capacity to capture intricate interactions between speech-independent factors and speech-driven movements. To achieve fine-grained control over speech-independent factors, some approaches [23, 25, 28] have proposed disentangling speech-driven and speech-independent movements in a single 3D face sequence. However, these methods tend to oversimplify or only

[*] Corresponding author.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3611775

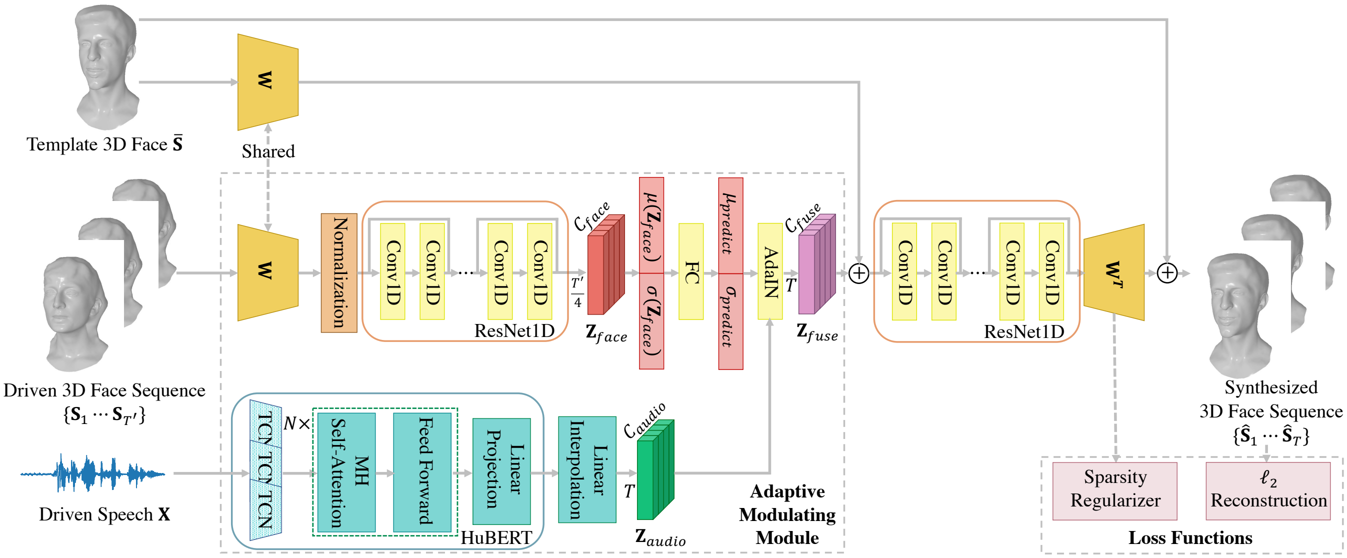


Figure 1: The overall framework of our method. The adaptive modulating module incorporates the composite nature of facial movements into the framework, while the sparsity regularizer interprets the regional nature of facial movements. The overall backbone is non-autoregressive, which enables efficient training and inference.

focus on local aspects of the composite nature, reducing the expressiveness of facial animations. Moreover, previous learning-based methods tend to overlook the regional nature of facial movements, while rule-based methods [7, 31, 39, 41] consider such nature but require extensive manual labor when animating unseen faces. To address the issues above, there is a critical need to develop a comprehensive method that captures a global understanding of the composite nature and considers the regional nature.

To tackle these challenges, we propose a novel speech-driven 3D face animation method that considers both facial movements' composite and regional natures. We introduce an adaptive modulating module to account for the composite nature. This module inputs the latent audio features and arbitrary 3D face sequence, extracting global-aware speech-independent representations and modulating the latent audio features according to the extracted representations. To accommodate the regional nature, we propose a sparsity regularizer, which, for each frame, enforces each facial feature element to focus on the local region of mesh vertices. Furthermore, we present a non-autoregressive backbone for translating audio to 3D facial movements. We apply the pretrained HuBERT model [14] to extract high-level audio features and adopt ResNet [13] with 1D convolution to serve as the motion decoder. The overall framework is shown in Figure 1. Our backbone both enables efficient inference and preserves high-frequency motion details.

To demonstrate the effectiveness of our framework, we conduct extensive experiments on the VOCA [5], MeshTalk [28], and BIWI [11] datasets. We evaluate our framework against several state-of-the-art approaches using various quantitative metrics, including the lip vertex error, face error, and dynamic time wrapping error. Moreover, we conduct a user study to evaluate the naturalness of facial movements and the synchronization between speech and animation. The experimental results show that our proposed framework outperforms existing methods in terms of both quantitative metrics and subjective evaluations, verifying that it is beneficial to consider both composite and regional natures in the task of speech-driven 3D face animation. The code is publicly available at https://github.com/wuhaozhe/audio2face_mm2023.

2 RELATED WORK

Speech-driven face animation has received significant attention in previous literature. Considerable research has focused on animating 2D faces [1, 4, 6, 9, 12, 15, 18, 19, 26, 29, 36, 44], while we concentrate on animating 3D faces in this work. In 3D facial animation, rule-based methods have been previously explored [7, 31, 39, 41]. These methods rely on breaking down facial movements into smaller units, such as visemes [24] and facial action units (FAUs) [8], and establishing mappings between speech and these units. These rule-based methods take into account the regional nature of facial movements, with both visemes and FAUs designed based on anatomical characteristics of the human face. As a result, they have achieved satisfactory results in synthesizing lip motions. However, these methods require extensive manual labor when animating unseen faces and their performance is limited in synthesizing speech-independent movements, such as facial expressions that are not directly related to speech.

With the advent of 4D face datasets [5, 11, 28, 38], various learning-based methods have emerged [17]. A few methods focus on driving one particular character [20, 45], while most methods work on driving different identities [10, 12, 19, 23, 25, 28, 30, 32, 37, 40, 42, 43]. In these methods, researchers have attempted to consider the composite nature. Some methods regard speech-independent factors as one-hot labels. For example, methods such as FaceFormer [10] and CodeTalker [40] treated the one-hot identity

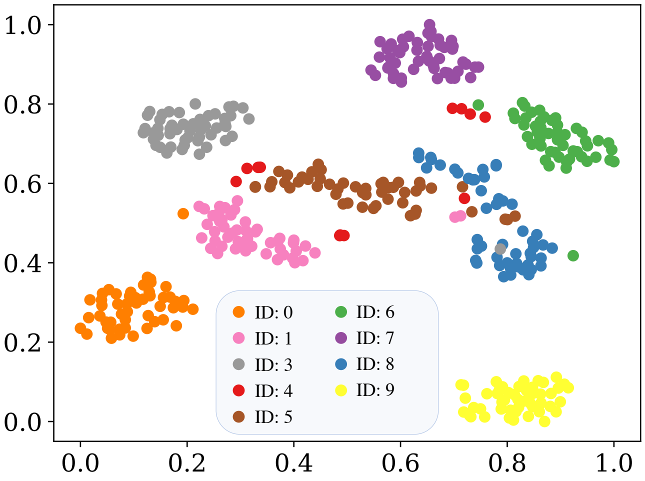


Figure 2: The t-SNE visualization of how speech-independent factors influence facial movement distributions. Each point represents the statistics of one 3D sequence. Points belonging to different identities are colorized with different colors.

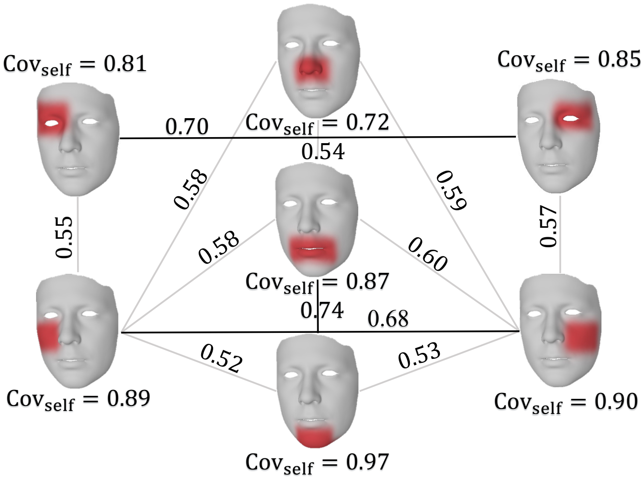

$\mathrm{Cov}_{\mathrm{self}} = 0.81$
$\mathrm{Cov}_{\mathrm{self}} = 0.72$
$\mathrm{Cov}_{\mathrm{self}} = 0.85$
0.70
0.54
0.58
0.59
0.55
0.58
0.60
0.57
$\mathrm{Cov}_{\mathrm{self}} = 0.87$
0.74
0.68
0.52
0.53
$\mathrm{Cov}_{\mathrm{self}} = 0.89$
$\mathrm{Cov}_{\mathrm{self}} = 0.90$
$\mathrm{Cov}_{\mathrm{self}} = 0.97$

Figure 3: The correlation graph of local facial regions. The local facial regions are colorized red. The $\mathrm{Cov}_{\mathrm{self}}$ denotes the self-correlation inside the local region, and the weight of the edge denotes the correlation between the two regions.

label as speech-independent factors and fused identity label with speech audio in the Transformer [35] decoder. These two methods synthesize discrete talking styles for each identity and enable interpolation between different styles. However, they are limited to synthesizing unseen talking styles, which poses a challenge for their practical applications. In the footsteps of the FaceFormer and CodeTalker methods, the Imitator [32] approach introduced a style adaptive motion decoder, which allows for fine-tuning on previously unseen styles. Nonetheless, this fine-tuning process incurs a significant computational cost and hampers fast generalization. In addition to identity labels, the SpaceX method [12] also incorporates one-hot emotion labels to generate expressive facial animations. Although these methods that utilize one-hot speech-independent labels can produce diverse facial movements, the representation granularity of speech-independent factors is often too coarse, which hinders their broader application.

To achieve fine-grained control over speech-independent factors, several methods [19, 23, 25, 28, 30, 37, 42] have proposed to disentangle speech-driven and speech-independent movements in a single 3D face sequence. Some of these methods achieve disentanglement in a supervised manner [19, 25, 28], such as the MeshTalk [28] method, which uses a cross-modality loss to disentangle the speech-driven and speech-independent facial movements. However, the assumption made by MeshTalk that speech-driven and speech-independent movements correspond respectively to the lower and upper parts of human faces may not hold for all facial expressions or movements. In contrast, the Emotional Video Portraits (EVP) [19] and EmoTalk [25] methods improve on MeshTalk by disentangling speech emotion and speech content through cross-reconstruction without making such assumptions. Nonetheless, these methods still require one sentence to be spoken with different emotions, which poses challenges for dataset collection. Some methods achieve disentanglement in an unsupervised manner [23, 30, 37, 42], the MemFace method [30] uses a latent memory dictionary to disentangle speech-independent factors. In contrast, the GeneFace [42] method incorporates uncertainty into the synthesis model to represent speech-independent movements statistically. In addition, some methods [23, 37] directly extract speech-independent representations from the reference video and fuse such representations with audio embeddings. To conclude, while the aforementioned methods have shown promising results in modeling speech-independent movements, they tend to oversimplify or only focus on the local aspects of the composite nature. The global understanding of the composite nature is lacking in these methods. Moreover, these learning-based methods neglect the regional nature of facial movements, further hindering the ability to synthesize realistic and vivid animations. These limitations highlight the need for a comprehensive approach that considers both facial movements' composite and regional natures for improved 3D facial animation synthesis.

3 OBSERVATIONS OF 3D FACE ANIMATION

In this section, we systematically investigate the impact of the composite and regional natures on the 3D face animations. To this end, we conduct data observations for each nature separately.

We explore the impact of the composite nature by visualizing how speech-independent factors influence the distribution of facial movements. Specifically, we calculate the standard deviation for each vertex of a 3D sequence along the temporal dimension, then reduce the standard deviations of all 3D mesh vertices to 2D using t-SNE [34], which we visualize as 2D points. Figure 2 demonstrates the visualization results. Different colors are used to plot points belonging to different identities. Our findings suggest that the speech-independent factor such as speaker identity significantly impacts the distribution of facial movements.

To investigate the impact of the regional nature on facial movements, we analyze motion correlation across different regions of the face. Motion correlation is used to quantify the degree of dependency between the movements of different facial regions. We first partition facial movements into small regions based on the facial mesh vertices to obtain the motion correlation. We then calculate the pairwise correlation coefficient along the spatial dimension between each pair of regions, resulting in a correlation matrix. Finally, we visualize this matrix as a connected graph, where edges with less than 0.5 correlation coefficient are removed. Figure 3 shows the correlation graph. The correlation graph reveals that the motion correlation of facial regions is not uniformly distributed, with some regions being more correlated than others. For example, the mouth and chin regions, which are responsible for similar expressions, display a higher correlation, while the upper and lower parts of the face have a lower correlation.

These findings have important implications for understanding the underlying mechanisms of 3D facial movements. In the next section, we propose a framework that comprehensively considers both composite and regional natures.

4 METHODOLOGY

As mentioned earlier, the composite and regional natures have a significant impact on facial movements, and it is crucial to take them into account when generating realistic and accurate 3D face animations. In this section, we detailedly elaborate on how we integrate these two natures into the speech-driven 3D face animation framework. Moreover, we introduce our non-autoregressive backbone, which preserves high-frequency details of facial animations and enables efficient inference.

4.1 Problem Formulation

Before introducing the overall framework, we first formulate the problem of audio-driven 3D face animation in the presence of speech-independent factors. This task takes three inputs: a template 3D face $\bar{\mathrm{S}}$ of the target person, the driven speech $\mathbf{X}$ with duration T, and the driven 3D face sequence $\{\mathrm{S}_1 \cdots \mathrm{S}_{T'}\}$. The objective is to synthesize a sequence of 3D face animations $\{\hat{\mathrm{S}}_1 \cdots \hat{\mathrm{S}}_T\}$. The synthesized animations have the same identity as $\bar{\mathrm{S}}$, are synchronized to the driven speech $\mathbf{X}$, and incorporate the speech-independent facial movements of $\{\mathrm{S}_1 \cdots \mathrm{S}_{T'}\}$.

It is worth noting that we do not synthesize the blendshape weights of 3D faces but rather directly synthesize the 3D face vertices. There are two main reasons for this choice. Firstly, blendshapes often result in a loss of high-frequency facial information, whereas the 3D face sequence preserves all facial details. By synthesizing 3D face sequences directly, our model can capture intricate facial movements and fine-grained nuances. Secondly, the blendshape weight is often limited by its uninterpretable definitions, whereas directly animating 3D faces has broader applications. Based on such a setting, all of the input 3D faces and the synthesized 3D faces have a shape of $N \times 3$, where N is the number of mesh vertices. For different datasets, the vertex number N is different.

4.2 Adaptive Modulating Module

To incorporate the composite nature in 3D facial animation synthesis, it is essential to effectively combine both speech-independent and speech-driven facial movements. To achieve this, we propose the adaptive modulating module, which plays a critical role in effectively blending these two types of movements. This module first extracts global-aware speech-independent representations that capture the facial movements not influenced by the speech signal. These representations are then used to modulate the latent audio features, allowing the model to dynamically adjust the contribution of speech-driven and speech-independent factors to each specific facial region. By utilizing the adaptive modulating module, our framework synthesizes more diverse and natural face animations.

Figure 1 shows the adaptive modulating module. The module extracts the speech-independent representations from $\{\mathrm{S}_1 \cdots \mathrm{S}_{T'}\}$. More specifically, for each 3D face S_i, we first normalize S_i by subtracting the mean face of $\{\mathrm{S}_1 \cdots \mathrm{S}_{T'}\}$. Formally:

$$\mathrm{Norm}(\mathrm{S}_i) = \mathrm{S}_i - \frac{\sum_{i=1}^{t} \mathrm{S}_i}{t}. \tag{1}$$

The goal of normalization is to extract solely the facial movement information while removing the identity information. Subsequently, we reduce $\mathrm{Norm}(\mathrm{S}_i)$ to a low-dimensional feature vector with embedding matrix $\mathbf{W}$. We then concatenate the embedded vector of the 3D faces into a sequence and input this sequence into a ResNet1D [13] encoder, thereby obtaining the latent face representations $\mathbf{Z}_{face}$. $\mathbf{Z}_{face}$ has a shape of $\frac{T'}{4} \times C_{face}$, where C_{face} is the channel number, and $\frac{T'}{4}$ accounts for the downsampled embedded face vectors along the temporal dimension. Afterward, we extract the speech-independent facial movements from $\mathbf{Z}_{face}$. Different from the previous methods [19, 25, 28] that leverages cross reconstruction loss to extract speech-independent factors, we extract the speech-independent information by simply calculating the mean $\mu(\cdot)$ and standard deviation $\sigma(\cdot)$ of $\mathbf{Z}_{face}$ along the temporal dimension. Remarkably, this simple approach provides an effective representation of speech-independent information, as it captures the overall statistical distribution of face animations while excluding the temporal information of $\mathbf{Z}_{face}$.

Having obtained $\mu(\mathbf{Z}_{face})$ and $\sigma(\mathbf{Z}_{face})$, we now blend them with input speech signals. We first feed the input speech to the pretrained audio model [14], yielding the latent audio feature $\mathbf{Z}_{audio}$ with a shape of $T \times C_{audio}$, where C_{audio} is the channel number of latent audio features. Afterwards, $\mu(\mathbf{Z}_{face})$ and $\sigma(\mathbf{Z}_{face})$ are used to modulate the mean and standard deviation of $\mathbf{Z}_{audio}$ on a global level. Specifically, we map $\mu(\mathbf{Z}_{face})$ and $\sigma(\mathbf{Z}_{face})$ to $\mu_{predict}$ and $\sigma_{predict}$ with a linear layer, where $\mu_{predict}$ and $\sigma_{predict}$ have the same channel number as $\mathbf{Z}_{audio}$. Finally, we adjust $\mathbf{Z}_{audio}$ with a similar manner as AdaIN [16]:

$$\mathbf{Z}_{fuse} = \sigma_{predict}\left(\frac{\mathbf{Z}_{audio} - \mu(\mathbf{Z}_{audio})}{\sigma(\mathbf{Z}_{audio})}\right) + \mu_{predict}. \tag{2}$$

The acquired $\mathbf{Z}_{fuse}$ contains both speech driven and speech independent information.

4.3 Sparsity Regularizer

When we linearly embed $\mathrm{Norm}(\mathrm{S}_i)$ to low-dimensional facial feature vectors with embedding matrix $\mathbf{W}$, it is necessary to consider the regional nature of the facial movements, or the resulting feature vectors may fail to capture the subtle details in local regions.

A simple and straightforward method for incorporating regionality is to divide the face into several regions according to face anatomy and apply a separate embedding matrix to each region. In this way, each region can be embedded independently and with greater detail. However, this approach can be labor-intensive as it requires the manual splitting of 3D faces into different regions, which is time-consuming and require expert knowledge. Alternatively, some previous methods [5] also initializes $\mathbf{W}$ from parametric 3D face model [22]. Such initialization does help the model to capture better facial details, but it can only synthesize 3D face mesh which has the same topology as the parametric 3D face model. When we switch the 3D face template, this method fails.

Our proposed approach aims to address the issue of regional nature by leveraging a novel and efficient strategy, which does not require manual labor and is applicable to different 3D face templates. To achieve this, we utilize a sparse regularization technique inspired by Lasso Regression [33]. In particular, we apply ℓ_1 regularization to the embedding matrix $\mathbf{W}$, which encourages several elements of the weight matrix to be close to zero, resulting in sparsity. The sparsity enables each element of the feature vector to focus on the local facial regions. Furthermore, this strategy also improves the interpretability of the learned weights and leads to better generalization capability.

4.4 Backbone

Designing an efficient and effective backbone is also crucial for the task of audio-driven 3D face animation. In this section, we illustrate how the backbone obtains the latent audio feature $\mathbf{Z}_{audio}$ and how the backbone generates $\{\hat{\mathrm{S}}_1 \cdots \hat{\mathrm{S}}_T\}$ from $\mathbf{Z}_{fuse}$ and $\bar{\mathrm{S}}$.

We utilize the pre-trained HuBERT model [14] for audio encoding. Notice that we have compared HuBERT, wav2vec 2.0 [3], Mel spectrogram, and DeepSpeech [2] features. Among these features, we observe that the HuBERT feature performs best. The HuBERT model is a self-supervised method for learning audio representations that achieves state-of-the-art performance on various downstream tasks. The model is designed to take raw audio waveforms as input and generate a sequence of high-level representations that capture various aspects of the audio signal. In our implementation, we extract the final layer of the HuBERT model and resample the output with the desired frame rate to obtain the latent audio feature $\mathbf{Z}_{audio}$. During the training process, we do not fix the HuBERT model as the previous method [10] does. Instead, we adopt a warm-up strategy. More specifically, at the start of training, we fix the HuBERT model and train the other sub-modules. Once the other sub-modules almost converge, we unfreeze the HuBERT model to allow it to fine-tune the task. Such a strategy has the advantage of preventing the scratch-initialized sub-modules from disturbing the pre-trained HuBERT model, which already contains useful and high-quality audio representations. With the warm-up strategy, we achieve faster convergence and better performance compared to simply freezing the pretrained audio model.

For the other sub-modules in our backbone, we extensively employ the ResNet1D [13] structure rather than the Transformer [35] structure. The ResNet1D conducts 1D convolution on the input feature vector sequence along the temporal dimension. It has the following three characteristics: (1) ResNet1D imposes a strong inductive bias on the model architecture, which aggregates information from temporal-adjacent frames. Such inductive bias lessens the required data for training. (2) ResNet1D has a strong capability for non-linear translation due to the stack of numerous convolution layers. (3) ResNet1D is a non-autoregressive and fully convolutional architecture, thereby it requires less computation cost and adapts to input with arbitrary size. These characteristics tally well with the task of speech-driven 3D face animation. Usually, the input speech and 3D facial animation have strict temporal correspondence, such property has lessened the requirement of building complex time dependencies. When mapping speech to 3D facial animation, it is sufficient to fuse temporal information from adjacent frames as the ResNet1D does. Moreover, the input speech and 3D facial animation are highly heterogeneous, therefore the powerful non-linear translation capacity of ResNet1D is in need for our task.

Based on the intuition above, we generate $\{\hat{\mathrm{S}}_1 \cdots \hat{\mathrm{S}}_T\}$ from $\mathbf{Z}_{fuse}$ and $\bar{\mathrm{S}}$ with the ResNet1D decoder. We first embed $\bar{\mathrm{S}}$ with the embedding matrix $\mathbf{W}$ mentioned in Section 4.3, and then add the embedded vector to each frame of $\mathbf{Z}_{fuse}$. Afterward, the ResNet1D decoder takes the added embedding as input, and outputs the predicted movement features $\mathbf{Z}_{pred}$ with shape $T \times C_{face}$. Based on $\mathbf{Z}_{pred}$, we synthesize the 3D face sequences with the following equation:

$$\hat{\mathrm{S}}_{\mathbf{i}} = \bar{\mathrm{S}} + \alpha \mathbf{W}^T \times \mathbf{Z}_{pred}[i], \tag{3}$$

where $\mathbf{W}^T$ is the transposed matrix of the embedding matrix $\mathbf{W}$. We scale the predicted movements with a coefficient α for faster convergence, α is set to 0.1 in our implementation. Notice that we have removed all of the downsampling layers in ResNet1D during the decoding process; all layers have a convolutional kernel with size 3 and stride 1. Such modification avoids synthesizing over-smooth animations and retains high-frequency details.

Overall, both the encoder and the decoder of our backbone are non-autoregressive, thus can be trained in parallel and run efficiently during inference. The non-autoregressive design also allows for flexible and variable-length input sequences, making our model applicable to various applications.

Training objectives. We simultaneously optimize the ℓ_2 loss and the sparsity regularization loss. The ℓ_2 loss calculates the distance between the synthesized 3D face sequences $\{\hat{\mathrm{S}}_1 \cdots \hat{\mathrm{S}}_T\}$ and the ground truth 3D face sequences $\{\mathrm{S}_1 \cdots \mathrm{S}_T\}$. The sparsity regularization loss minimizes the ℓ_1 norm of $\mathbf{W}$. Formally:

$$\mathcal{L} = \mathcal{L}_{\ell_2} + \beta \mathcal{L}_{\text{reg}} = \sum_{i=1}^{T} ||\hat{\mathrm{S}}_i - \mathrm{S}_i||_2 + \beta ||\mathbf{W}||_1. \tag{4}$$

4.5 Implementation Details

For the HuBERT model, we adopt the HuBERT-large configuration with 24 Transformer layers. For the ResNet1D model, we adopt the ResNet18 configuration. The mesh embedding matrix $\mathbf{W}$ has a shape of $3N \times 256$, where N is the number of mesh vertices. Different datasets have different numbers of mesh vertices. During training, we leverage the Adam optimizer [21] with learning rate of 10^{-4}. The weight decay of the Adam optimizer is set to 0 because it contradicts our sparsity regularizer. We train for 120 epochs with a mini-batch size of 8 samples. In the implementation, the scaling coefficient β of the regularization loss is set to 10^{-4}. We fix the HuBERT model at the first ten training epochs and unfreeze it at the subsequent epochs. We leverage one RTX3090 GPU for training. The training process takes less than one hour.

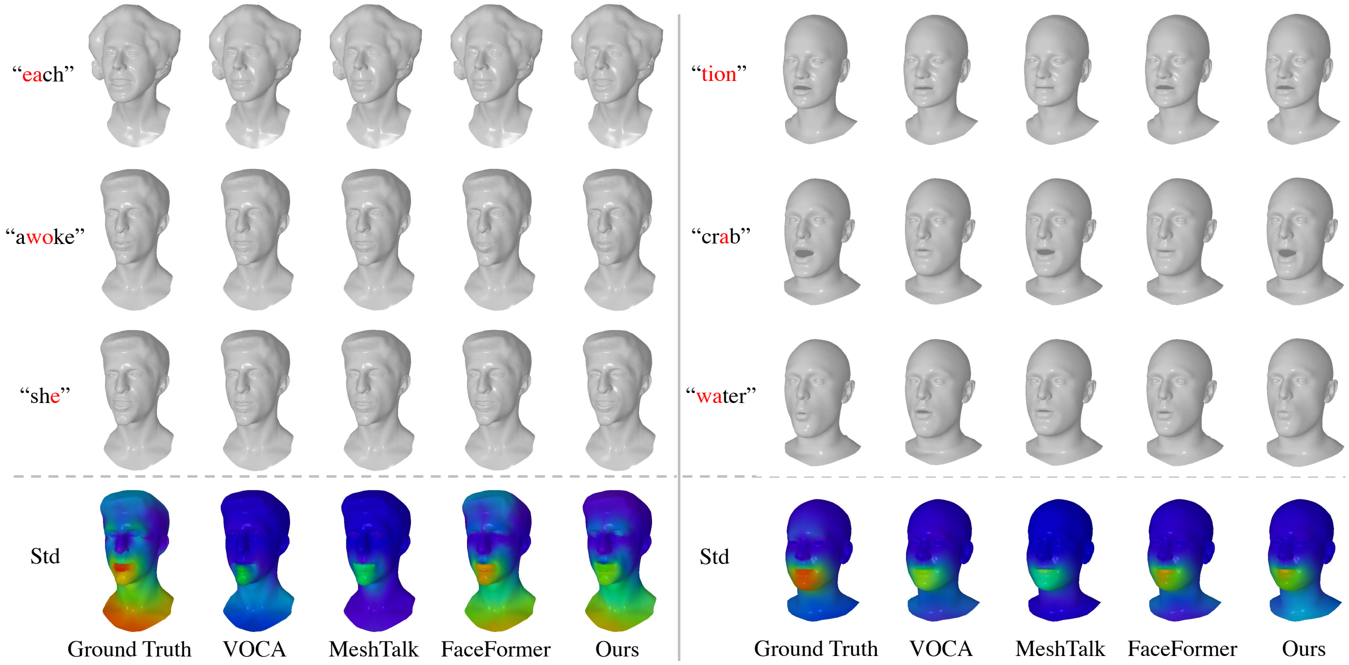


Figure 4: Qualitative comparison with baseline methods on MeshTalk dataset (left) and VOCASET (right). The first three rows show the facial animations when speaking different phonemes. The bottom row shows the standard deviation of facial animations; red denotes a large standard deviation, while blue denotes a smaller one.

5 EXPERIMENTS

5.1 Experimental Settings

VOCASET dataset [5]. VOCASET contains 480 3D face sequences obtained from 12 individuals. The 3D face mesh of each sequence adopts the template of the FLAME face model [22] with 5023 vertices. We adopt the same training, validation, and testing splits for fair comparisons as VOCA [5]. Eight individuals are selected for training, two individuals are selected for validation, and two individuals are selected for testing.

BIWI dataset [11]. The BIWI dataset contains expressive emotions. Each mesh of the BIWI dataset contains 23370 vertices. We adopt the same evaluation protocol as FaceFormer [10] on the BIWI dataset. More specifically, we exclude the neutral sentences from BIWI during evaluation and select only the emotional sentences. The training set has 192 sentences, the validation set has 24 sentences, and the testing set has 32 sentences.

MeshTalk dataset [28]. The MeskTalk dataset is not fully opensourced. Among all of the 250 individuals in the dataset, only the 3D face animations of 13 individuals are publicly available. Each mesh of the MeskTalk dataset contains 6172 vertices. Among the 13 individuals, we selected nine individuals for training, four for validation, and two for testing. The testing and validation sets overlap partially in individuals.

Baseline Methods. We conducted a comparative evaluation of our proposed framework with three state-of-the-art methods: VOCA [5], MeshTalk [28], and FaceFormer [10]. While VOCA and FaceFormer methods are conditioned on the identity label and driven speech, the MeshTalk method is conditioned on the 3D face sequence and driven speech. To ensure a fair comparison between the baseline and our proposed methods, we adopted the evaluation protocols of the respective methods. For the evaluation of VOCA and FaceFormer methods, we synthesized 3D face sequences based on test speech and training identities following the evaluation protocol of FaceFormer. For the evaluation of the MeshTalk method and our approach, we follow the evaluation protocol of MeshTalk, which obtains 3D face sequences from test speech and test 3D face sequences. During the evaluation process, we took great care to ensure no information leakage between the synthesized 3D face sequences and the testing 3D face sequences. The evaluation metrics were computed between the synthesized 3D face sequences and the testing 3D face sequences.

5.2 Quantitative Evaluations

We quantitatively evaluate the synchronization between the driven audio and the synthesized 3D face animations. To evaluate lip synchronization, we adopted two metrics: maximal lip vertex error (L^{lip}_{max}) and average lip vertex error (L^{lip}_{mean}). L^{lip}_{max} firstly computes the maximum Euclidean distance of lip region vertices between the synthesized and the ground truth 3D face and then averages the error among frames. L^{lip}_{mean} calculates the average distance between lip region vertices. To evaluate the synchronization of speech-independent movements, we utilize the following metrics: average upper face error (L^{upper}_{mean}) and average face error (L^{face}_{mean}). L^{upper}_{mean} and L^{face}_{mean} respectively calculate the average Euclidean distance of the upper and the whole face. Additionally, we calculate face dynamic time wrapping error (F-DTW), which compares the temporal dynamics of the synthesized and ground truth 3D face sequences. F-DTW measures the similarity of two temporal sequences by finding an optimal warping path to align the sequences in time.

Table 1 presents the comparison results among the methods. Notably, the MeshTalk method shows suboptimal performance due to its original implementation's heavy reliance on large-scale training data, which is not available in our experimental setup. As a result, the MeshTalk method exhibits inadequate generalization capacity when the training data only comprises around ten individuals. In contrast, the FaceFormer and our methods have better generalization capacity due to the incorporation of pretrained audio models [3, 14]. Furthermore, our method outperforms the FaceFormer method by a large margin in terms of lip synchronization and speech-independent synchronization.

In addition, our method is also computationally efficient due to the design of non-autoregressive architecture. Our method takes 0.007 seconds to synthesize 1-second 3D face sequences during inference, while the FaceFormer method takes 0.1 seconds. The efficiency of our method makes it practical for real-time applications such as video conferencing, telepresence, and gaming.

5.3 Qualitative Evaluations

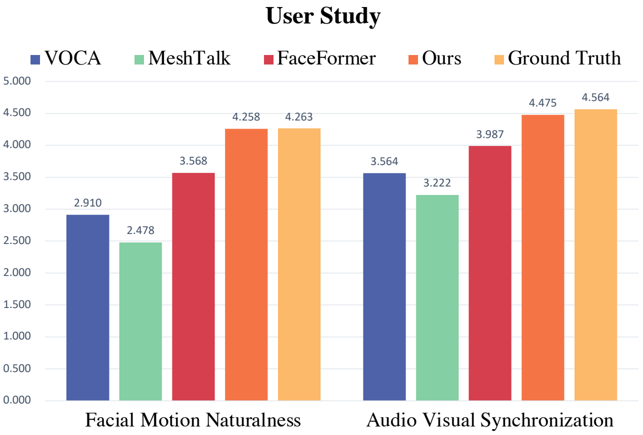


Figure 5: User study results. The average scores of facial motion naturalness and audio-visual synchronization are reported. Higher scores denote better results.

We qualitatively compare our method with baseline methods in Figure 4. The first three rows give the synthesized 3D faces for different input speeches. For fair comparisons, all methods and input speeches are required to utilize the same talking style. Our method generates more realistic and vivid 3D facial movements with better lip synchronization than the baseline methods. Mainly, our method exhibits more significant mouth opening and closing movements for pronouncing /b/ and /p/ and more evident pouting movements for pronouncing /w/ and /v/. Meanwhile, the baseline methods affect jaw flapping, resulting in unnatural facial movements.

The bottom row demonstrates that our and FaceFormer methods synthesize diversified speech-independent facial movements. We generate a heat map to visualize the intensity of facial movements, with vertices that exhibit a larger distance of motion appearing in red. In comparison, those with a shorter distance appear in blue. We observe that our and FaceFormer methods synthesize more intense movements for the lip region and the other facial regions. In contrast, the VOCA and MeshTalk methods tend to synthesize over smooth movements with little variation in intensity. By generating more diverse and intense facial movements, our method can create more vivid and expressive 3D faces, enhancing the overall realism of the synthesized output.

Additionally, we conduct user studies to evaluate the naturalness of facial motion and audio-visual synchronization. Specifically, we randomly select eight audio samples to drive the 3D face animations. We synthesize the 3D face animations with the same talking style as the input audio for each method. We invite 14 participants to rate the facial motion naturalness and audio-visual synchronization. The participants are asked to provide ratings on two aspects: (1) whether the facial motion appears natural and (2) whether the audio and 3D face animation are synchronized properly. Participants rate the mean opinion score (MOS) [27] using a 1-5 scale, with higher scores indicating better results. Figure 5 reports the results of user studies. We received feedback from several participants indicating that VOCA and MeshTalk methods exhibit unnatural facial movements due to the jaw-flapping effect. Compared with the FaceFormer method, our method synthesizes facial movements with more subtle details. Overall, these results highlight the effectiveness of our proposed method in generating high-quality 3D facial animations.

5.4 Ablation Study

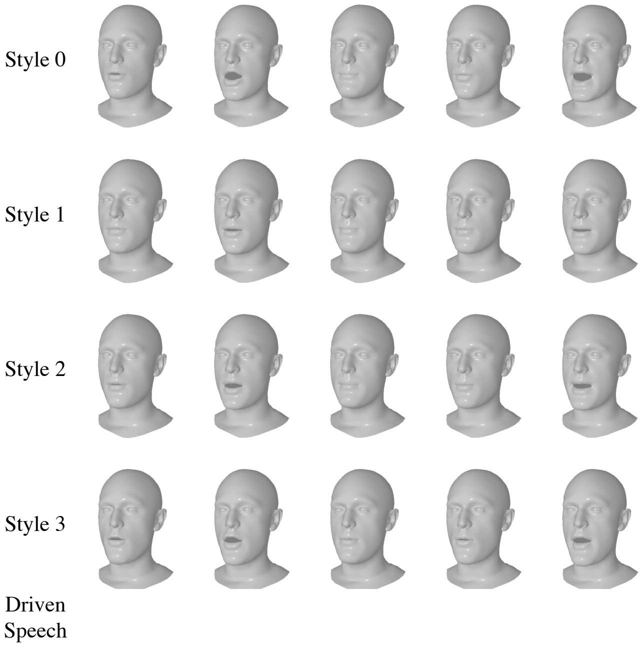


Figure 6: Visualizations of stylized 3D facial animations driven by input speech.

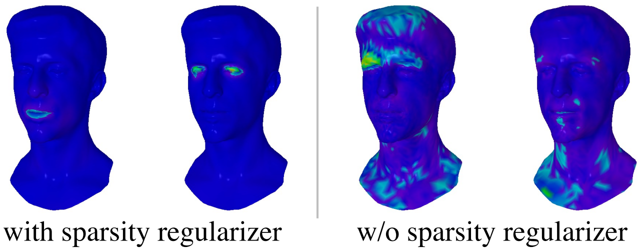
We conducted ablation studies to evaluate the effectiveness of considering composite and regional natures in synthesizing speech-driven 3D face animations. Specifically, we performed experiments

Table 1: Comparison with state-of-the-art methods, lower denotes better for all metrics. Our method outperforms baseline methods in terms of both speech-driven movements and speech-independent movements. Notice that the scaling of metrics across different datasets is inconsistent due to variations in the scales of the original data.

Dataset	VOCASET [5]					MeshTalk dataset [28]					BIWI dataset [11]				
Method	L^{lip}_{mean}	L^{lip}_{max}	L^{upper}_{mean}	L^{face}_{mean}	F-DTW	L^{lip}_{mean}	L^{lip}_{max}	L^{upper}_{mean}	L^{face}_{mean}	F-DTW	L^{lip}_{mean}	L^{lip}_{max}	L^{upper}_{mean}	L^{face}_{mean}	F-DTW
VOCA [5]	0.00324	0.00630	0.00054	0.00091	0.207	2.778	4.968	0.717	1.323	135.5	0.0235	0.0429	0.0089	0.0136	1.948
MeshTalk [28]	0.00350	0.00640	0.00055	0.00092	0.210	2.516	4.556	0.776	1.268	129.3	0.0227	0.0424	0.0082	0.0126	1.779
FaceFormer [10]	0.00212	**0.00438**	0.00046	0.00077	**0.091**	2.206	3.885	0.711	1.210	123.9	0.0230	0.0402	0.0092	0.0143	2.047
Ours w/o Composite	0.00171	0.00470	**0.00041**	0.00064	0.145	1.728	3.474	0.631	0.938	96.6	0.0186	0.0381	0.0071	0.0120	1.491
Ours w/o Regional	0.00166	0.00451	**0.00041**	**0.00063**	0.142	1.690	3.416	0.632	**0.927**	**95.6**	0.0175	0.0366	0.0069	0.0107	1.376
Ours	**0.00161**	0.00447	0.00042	0.00065	0.147	**1.659**	**3.382**	**0.621**	0.930	96.2	**0.0170**	**0.0353**	**0.0068**	**0.0105**	**1.330**

by selectively removing the composite or regional nature from our model architecture and synthesizing speech-driven 3D face animations using the modified models. The quantitative results of these experiments are reported in Table 1.

The results of the MeshTalk and BIWI datasets reveal that removing the regional nature from our model has a significant negative impact on the performance of lip synchronization. This is because the regional nature allowed our model to focus on specific details of facial movements around the lips, which are crucial for accurately synchronizing lip movements with speech. On the other hand, when we remove the composite nature from our model, we observed a decline in performance for synthesizing speech-independent facial movements. This is because the composite nature enabled our model to capture global patterns and general trends in facial movements that are independent of speech. It is worth noting that the VOCASET dataset has minimal speech-independent facial movements, and hence, the performance inside the ablation study was similar. Our findings suggest that composite and regional natures are important for synthesizing speech-driven 3D face animations.


with sparsity regularizer

w/o sparsity regularizer

Figure 7: The activated region for each element of the motion features. Blue regions indicate no influence, while red regions denote high influence, and green regions represent a transitional influence between blue and red.

We have also created visualizations to showcase the effectiveness of our method in synthesizing composite and regional facial movements. Figure 6 shows that our method successfully generates diverse talking styles for single-driven speech input, highlighting the ability to effectively capture the nuances and variations in facial expressions characteristic of different speaking styles. To further demonstrate the effectiveness of our method, we also draw Figure 7 to illustrate the impact of our proposed sparsity regularizer, which enforces each facial feature element to focus on the local region of mesh vertices. By incorporating this sparsity regularizer, our method can identify and extract interpretable regions for synthesizing facial movements, leading to more natural and accurate results. Meanwhile, when the sparsity regularizer is removed, the activated regions of motion features spread across the face.

Overall, our ablation study provides empirical evidence for the effectiveness of considering both composite and regional natures in synthesizing speech-driven 3D face animations. By combining these two natures, our model can achieve superior performance.

6 CONCLUSION

This paper emphasizes the importance of considering composite and regional natures in speech-driven 3D face animation. We conducted extensive observations demonstrating that these natures are prevalent in 3D facial movements. Our proposed comprehensive 3D face animation framework incorporates both of these natures. To handle the composite nature, we introduced an adaptive modulating module that extracts speech-independent information from arbitrary 3D face sequences and fuses this information with the driving audio. To address the regional nature, we proposed a sparsity regularizer that enforces each element of the motion feature to focus on local regions of 3D faces. Furthermore, we designed an efficient non-autoregressive backbone for mapping audio and 3D facial movements. Our backbone is built on a pretrained HuBERT model and a ResNet1D network, which preserves high-frequency details of facial movements. During implementation, our backbone synthesizes one second of facial animations with 30 fps in only 0.007 seconds. Extensive experiments demonstrate that our proposed framework outperforms baseline methods both quantitatively and qualitatively, with significantly reduced computational cost.

7 ACKNOWLEDGEMENTS

This work is supported by the National Key R&D Program of China under Grant No. 2020AAA0108600, the State Key Program of the National Natural Science Foundation of China (NSFC) (No.61831022), and supported in part by the Natural Science Foundation of China under Grant No.62222606 and 62076238.

REFERENCES

[1] Mohammed M Alghamdi, He Wang, Andrew J Bulpitt, and David C Hogg. 2022. Talking Head from Speech Audio using a Pre-trained Image Generator. In *ACM International Conference on Multimedia*. 5228–5236.

[2] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*. PMLR, 173–182.

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*. 12449–12460.

[4] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2018. Lip movements generation at a glance. In *European Conference on Computer Vision*. 520–535.

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

[6] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. 2020. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *European Conference on Computer Vision*. 408–424.

[7] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. 2016. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics* 35 (2016), 1–11.

[8] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

[9] Bo Fan, Lijuan Wang, Frank K Soong, and Lei Xie. 2015. Photo-real talking head with deep bidirectional LSTM. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 4884–4888.

[10] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. 2022. FaceFormer: Speech-Driven 3D Facial Animation with Transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18770–18780.

[11] Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. 2010. A 3-D audio-visual corpus of affective communication. *IEEE Transactions on Multimedia* 12 (2010), 591–598.

[12] Siddharth Gururani, Arun Mallya, Ting-Chun Wang, Rafael Valle, and Ming-Yu Liu. 2022. SPACEx: Speech-driven Portrait Animation with Controllable Expression. *arXiv preprint arXiv:2211.09809* (2022).

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 770–778.

[14] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.

[15] Ricong Huang, Weizhi Zhong, and Guanbin Li. 2022. Audio-driven Talking Head Generation with Transformer and 3D Morphable Model. In *ACM International Conference on Multimedia*. 7035–7039.

[16] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE/CVF International Conference on Computer Vision*. 1501–1510.

[17] Jewoong Hwang and Kyoungju Park. 2022. Audio-driven Facial Animation: A Survey. In *IEEE International Conference on Information and Communication Technology Convergence*. 614–617.

[18] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. 2022. EAMM: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH Conference*. 1–10.

[19] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. 2021. Audio-driven emotional video portraits. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14080–14089.

[20] 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* 36, 4 (2017), 1–12.

[21] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

[22] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics* 36, 6 (2017), 194:1–194:17.

[23] Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and Xin Yu. 2023. StyleTalk: One-shot Talking Head Generation with Controllable Speaking Styles. *arXiv preprint arXiv:2301.01081* (2023).

[24] Wesley Mattheyses and Werner Verhelst. 2015. Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication* 66 (2015), 182–217.

[25] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. 2023. EmoTalk: Speech-driven emotional disentanglement for 3D face animation. *arXiv preprint arXiv:2303.11089* (2023).

[26] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM International Conference on Multimedia*. 484–492.

[27] ITUT Recommendation. 2006. Vocabulary for performance and quality of service. *International Telecommunications Union—Radiocommunication* (2006).

[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 *IEEE/CVF International Conference on Computer Vision*. 1173–1182.

[29] Sanjana Sinha, Sandika Biswas, Ravindra Yadav, and Brojeshwar Bhowmick. 2022. Emotion-Controllable Generalized Talking Face Generation. In *International Joint Conference on Artificial Intelligence*. 1320–1327.

[30] Anni Tang, Tianyu He, Xu Tan, Jun Ling, Runnan Li, Sheng Zhao, Li Song, and Jiang Bian. 2022. Memories are One-to-Many Mapping Alleviators in Talking Face Generation. *arXiv preprint arXiv:2212.05005* (2022).

[31] Sarah L Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. 2012. Dynamic units of visual speech. In *ACM SIGGRAPH/Eurographics Conference on Computer Animation*. 275–284.

[32] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. 2022. Imitator: Personalized Speech-driven 3D Facial Animation. *arXiv preprint arXiv:2301.00023* (2022).

[33] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288.

[34] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605.

[35] 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), 5998–6008.

[36] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2020. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision* 128, 5 (2020), 1398–1413.

[37] Haozhe Wu, Jia Jia, Haoyu Wang, Yishun Dou, Chao Duan, and Qingshan Deng. 2021. Imitating Arbitrary Talking Style for Realistic Audio-Driven Talking Face Synthesis. In *ACM International Conference on Multimedia*. 1478–1486.

[38] Haozhe Wu, Jia Jia, Junliang Xing, Hongwei Xu, Xiangyuan Wang, and Jelo Wang. 2023. MMFace4D: A Large-Scale Multi-Modal 4D Face Dataset for Audio-Driven 3D Face Animation. *arXiv preprint arXiv:2303.09797* (2023).

[39] Zhiyong Wu, Shen Zhang, Lianhong Cai, and Helen M Meng. 2006. Real-time synthesis of Chinese visual speech and facial expressions using MPEG-4 FAP features in a three-dimensional avatar.. In *Annual Conference of the International Speech Communication Association*. 1802–1805.

[40] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[41] Yuyu Xu, Andrew W Feng, Stacy Marsella, and Ari Shapiro. 2013. A practical and configurable lip sync method for games. In *Motion on Games*. 131–140.

[42] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, Jinzheng He, and Zhou Zhao. 2023. GeneFace: Generalized and High-Fidelity Audio-Driven 3D Talking Face Synthesis. *arXiv preprint arXiv:2301.13430* (2023).

[43] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. 2020. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv e-prints* (2020), arXiv–2002.

[44] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI Conference on Artificial Intelligence*. 9299–9306.

[45] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. 2018. Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics* 37, 4 (2018), 1–10.