Speech Driven 3D Facial Animation Based on Blendshapes Parameters

Written by Students from Informatics Class[1,2,3] and AI Class[4,5]
Yixiang Zhuang[1], Chenkang He[2], Liangliang Chen[3], Yanzhe Zhao[4], Jiyao Chen[5]
[1]30920231154373, [2]30920231154347, [3]30920221154263
[4]30920231154370, [5]30920231154341

Abstract

The generation of 3D facial animations driven by speech faces significant challenges as it requires learning a many-to-many mapping between speech and the corresponding natural facial motion. Previous research has mainly focused on directly predicting 3D human facial motions from speech, whose the ground truth are coordinates. However, in this paper, we change the motion representation. Our approach aims to learn a concise set of blendshapes from speech. By combining the generated blendshapes with 3DMM(3D Morphable Model), we can generate more diverse faces. Compared to using only facial vertices to represent facial motion, using the blendshapes parameter as a prior can better model real facial motion. The loss is expressed as a weighted combination of blendshapes and vertices errors. To cope with the data scarcity issue, we use the self-supervised pre-trained speech representations as audio encoder. To address the issue of lip tremble, we integrate the transformer, which is well-suited for handling long contexts. This integration helps mitigate the lip tremble problem effectively. In our study, we conducted experiments on self-collected news broadcast datasets. The results of extensive experiments demonstrate that our approach can generate vivid facial animations while reducing computation.

Introduction

3D facial animation has been an active research topic for decades, as attributed to its broad applications in virtual reality, film production, and games. The objective of realistic speech-driven 3D facial animation is to automatically animate vivid facial expressions of a 3D avatar based on a given speech signal. The high correlation between speech and facial gestures (especially lip movements) makes it possible to drive the facial animation with a speech signal. Early attempts mainly focused on establishing complex mapping rules between phonemes and their visual counterpart, which typically had limited performance (Taylor et al. 2017a; Xu et al. 2013a). With the progress of deep learning, recent speech driven facial animation technology has rapidly developed. However, it still remains challenging to generate human-like motions.

In speech-driven 3D facial animation, most 3D mesh-based works use short audio windows as input, which may lead to blurriness in facial expression changes. Also, as a many-to-many mapping problem, speech-driven facial animation generally has multiple plausible outputs for every input. Such ambiguity tends to cause over-smoothed results. obviously, to achieve realistic animation of the entire face, a longer audio background is required. Although MeshTalk (Richard et al. 2022) considers longer audio contexts by modeling audio sequences, in the case of data scarcity, using Mel spectral audio feature training models cannot synthesize accurate lip movements. Collecting 3D motion capture data is also quite expensive and time-consuming. Regardless, person-specific approaches often achieve decent facial motions due to relatively consistent speaking styles, but are less scalable for general applications. Recently, VOCA (Cudeiro et al. 2019) has extended these methods to different identities, however, they usually exhibit indistinct upper facial motion. This is because VOCA formulates speech driven face motion mapping as a regression task, which encourages average motion, especially on upper surfaces that are only weakly or even uncorrelated with the speech signal. To reduce uncertainty, FaceFormer (Fan et al. 2022a) exploits long-term audio context via a transformer-based model and synthesizes sequential motion in an autoregressive manner. Although it offers significant improvements in performance, it still suffers from the disadvantages of one-to-one mapping and lacks subtle high-frequency motion.

We get inspiration from 3D Face Morphable Model (3DMM) (Yang et al. 2020a), where general facial expressions are represented in a low-dimensional space. Accordingly, we propose to formulate speech-driven facial animation as a task of learning facial expression parameters. The facial expression parameters also called blendshapes parameters. The network predicts blendshapes parameters and then uses a 3DMM model to fit and generate facial coordinates. We use blendshapes as a prior representation and calculate loss together with the fitted facial coordinates. Compared to using only facial vertices to represent facial motion, using the blendshapes parameter as a prior can better model real facial motion.

To address the issues of data scarcity and contextual information, while also reducing computational complexity to a certain extent, We propose a blendshapes parameters based temporary autoregressive model. Firstly, it effectively utilizes the self-supervised pre-trained speech representations

to handle the data scarcity issue. Secondly, the model architecture is based on transformers, which are widely recognized as effective in solving contextual information. Finally, we use the blendshapes parameter to represent facial motion, which is a set of 52 values. Based on the blendshapes parameter to synthesize facial animation, it can reduce computation to a certain extent.

Transformer (Vaswani et al. 2023) achieves remarkable performance in both natural language processing and computer vision tasks. Transformers can better capture long-range context dependencies compared to RNN-based models. Transformer's success is primarily due to its design incorporating a self-attention mechanism that effectively models both short-term and long-term relationships by paying explicit attention to all parts of the representation. The direct application of a standard transformer architecture to audio sequences tends to underperform in the task of speech-driven 3D facial animation, due to transformers are inherently data-hungry, necessitating large datasets for effective training.

Given the scarcity of 3D audio-visual data, we propose leveraging the self-supervised pre-trained speech model, wav2vec 2.0 (Baevski et al. 2020a). This model has been trained on a large-scale corpus of unlabeled speech, enabling it to learn rich phoneme information. Despite the limited coverage of phonemes in the available 3D audio-visual data, we anticipate that the pre-trained speech representations from wav2vec 2.0 can significantly enhance the performance of speech-driven 3D facial animation tasks, even in data-limited scenarios.

At the same time, in order to better characterize the face movement, we use the blendshapes parameter. We leverage blendshape parameters as prior information, employing 3D Morphable Model (3DMM) technology to align these parameters with facial vertex coordinates. Ultimately, we jointly compute the loss using both the blendshape parameters and the facial vertex coordinates, optimizing the entire task of voice-driven facial movement.

The main contributions of our work are as follows:

- **We propose a blendshapes parameters based temporary autoregressive model for speech driven facial animation.** It achieves highly realistic and temporally stable animation of the entire face including both the upper face and the lower face.
- **Loss function.** The loss function is weighted by the blendshapes parameter and the face vertex coordinate error.
- **Extensive experiments to assess the quality of synthesized face motions.** The results demonstrate that the model performs well in realistic facial animation and lip synchronization on our 3D dataset.

Related Work

Speech-driven 3D Facial Animation

Computer facial animation, a significant field in computer vision, has seen a notable increase in interest over recent years. Among its subfields, speech-driven facial animation, which aims to animate a virtual face in sync with a provided speech sequence, stands out. A substantial body of research focuses on 2D face animation (Alghamdi et al. 2022; Chen et al. 2020, 2018). However, in this study, we delve into the animation of 3D models. The approaches to this process can typically be classified into two categories: linguistics-based and deep learning based methods.

Linguistics-based methods. Linguistics-based methods are often extensively used, establishing complicated mapping rules between phonemes and their corresponding visual elements, namely, visemes. One such method is the dominance function (Massaro et al. 2001), which is designed to ascertain the effect of phonemes on facial animation control parameters. Additionally, Xu et al. (Xu et al. 2013b) determined animation curves within a devised canonical set of visemes to facilitate synchronised mouth movements. There are also some methods considering the many-to-many mapping between phonemes and visemes, as demonstrated in the JALI (Edwards et al. 2016). The JALI methodology divides mouth movements into lip and jaw rig animation. In doing so, JALI manages to deliver impressive co-articulation results. However, despite these procedures providing explicit control over the animation, they are relatively complex and lack a systematic approach in animating the entire face.

Deep learning based methods. Recently, Taylor et al. (Taylor et al. 2017b) introduced a deep learning-based model that employs a sliding window technique on the transcribed phoneme sequences input. VisemeNet (Zhou et al. 2018) utilized a sophisticated three-stage Long Short-Term Memory (LSTM) network to forecast the animation curve for a lower-face lip model.

We review the most relevant work here more specifically, as they have the same setup as this work, i.e. training on high-resolution paired audio mesh data and animating the entire facial mesh independently in vertex space. They are also based on deep learning. MeshTalk (Richard et al. 2022) successfully separates facial information that is correlated with audio from uncorrelated data, utilizing a categorical latent space. However, its latent space, although effective, doesn't offer optimal expressivity, often causing unstable animation quality in situations with scarce data. On the other hand, VOCA (Cudeiro et al. 2019), through the application of robust audio feature extraction models, is able to generate various styles of facial animation. Furthermore, FaceFormer (Fan et al. 2022a) brings in the concept of a longer-term audio context with a transformer, thereby producing temporally stable animations. Nevertheless, both VOCA and FaceFormer encounter an over-smoothing issue, which could be attributed to their direct regression approach for facial motion within the complex and widely varying audio-visual mapping domain, characterized by significant uncertainty and ambiguity.

3D Morphable Model

3DMM is a statistical model which transforms the shape and texture of the faces into a vector space representation. As 3DMM inherently contains the explicit correspondences from model to model, it is widely used in model fitting, face synthesis, image manipulations, etc. The recent research on

3DMM can be generally divided into two directions. The first direction is to separate the parametric space to multiple dimensions like identity, expression and visemes, so that the model could be controlled by these attributes separately. The models in expression dimension could be further transformed to a set of blendshapes (Li, Weise, and Pauly 2010), which can be rigged to generate individual-specific animation. Another direction is to enhance the representation power of 3DMM by using deep neural network to present 3DMM bases. The 3DMM model used in this article is FaceScape (Yang et al. 2020a), which is a bilinear model. The final face is equal to the base model times the expression parameters times the identity parameters.

The Proposed Framework

Overview

We cast speech-driven 3D facial animation into a sequence-to-sequence (seq2seq) learning framework and propose a novel seq2seq network architecture (Fig. 1). Our network takes audio, personal style, and previous facial movement sequence as input to predict the next frame's facial movement. In our framework (Fig. 1), the encoder first converts the audio into speech expression, and the style embedding layer encodes the speaker's vocal style into a set of learnable embeddings. The cross-modal decoder predicts the next frame's blendshape parameters based on past blendshape parameters, personal style, and audio features. Finally, the transform module converts the blendshapes into a Facescape head. In the following sections, we will elaborate on each component of our network architecture in detail.

Speech Encoder

Our encoder adopts the architecture of the state-of-the-art self-supervised pre-trained speech model, wav2vec 2.0 (Baevski et al. 2020b) The encoder comprises an audio feature extractor and a multi-layer transformer encoder. The audio feature extractor, composed of multiple temporal convolution layers (TCN), converts the raw waveform input into feature vectors. The transformer encoder stacks multi-head self-attention and feedforward layers, thereby transforming the audio features into contextualized speech embeddings. A quantization module discretizes the convolution outputs into a finite set of speech units. We leverage the context surrounding a masked time step to identify the true quantized speech unit by solving a contrastive task. To initialize our encoder (Fig. 1), we employ the pre-trained wav2vec 2.0 weights.

Cross-modal Decoder

The Cross-modal Decoder takes past blendshape parameters, personal style and audio features as input and autoregressively predicts the blendshape parameters of the next frame. The Cross-modal Decoder contains blendshape encoder, transformer decoder and blendshape decoder. The transformer decoder is equipped with the causal self- attention to learn the dependencies between each frame in the context of the past blendshape parameters, and the cross-modal attention to align the audio and motion modalities. The newly predicted parameters $\hat{\mathbf{b}}_t$ is used to update the past parameters as $\hat{\mathbf{B}}_{1:t}$ as preparation for the next frame prediction. The formula is as follows:

$$\hat{\mathbf{b}}_t = \mathbf{D}_{cross-modal}(A_{1:T}, \hat{\mathbf{B}}_{1:t-1}, style) \tag{1}$$

Transform Module

The Transform Module converts blendshapes, which is a way of deforming a mesh by interpolating between different shapes, to a Facescape head, which is a 3D head model that captures variations in identity, expression. This transform module enables our model to transfer facial expressions across different virtual characters quickly. This module is essentially a bilinear model from Facescape(Yang et al. 2020b). New face shape can be generated given the identity parameter $\mathbf{w}_{id}$ and expression parameter $\mathbf{w}_{exp}$ as:

$$V = \mathbf{C}_r \times \mathbf{W}_{id} \times \mathbf{W}_{exp} \tag{2}$$

where $\mathbf{C}_r \in \mathbb{R}^{78834\times52\times50}$ is a fixed core. $\mathbf{W}_{id} \in \mathbb{R}^{50}$ represents the identity parameter, which is fixed in the experiment. $\mathbf{W}_{exp} \in \mathbb{R}^{52}$ represents expression parameters and is the prediction of the network. $\mathbf{V} \in \mathbb{R}^{78834}$ represents 26,278 facial vertices.

Loss function

To train our neural network, we employ a loss function that comprises two distinct components: blendshapes loss and vertex loss. The overall function is given by:

$$L = \lambda_1 L_{blendshapes} + \lambda_2 L_{vertex} \tag{3}$$

where $\lambda_1 = 1E-2$,$\lambda_2 = 1.0$ in all of our experiments. We provide a detailed explanation of each of these components below.

Blendshapes loss. Given input audio A, the encoder extracts audio features and then sends them to the decoder to predict the corresponding blendshape parameters. Finally, Mean Squared Error (MSE) is applied between the predicted blendshape parameters $\hat{\mathbf{B}}_t = (\hat{\mathbf{b}}_1, ..., \hat{\mathbf{b}}_T)$ and the ground truth $\mathbf{B}_t = (\mathbf{b}_1, ..., \mathbf{b}_T)$. The formula is as follows:

$$L_{blendshapes} = \sum_{t=1}^{T} \parallel \hat{b}_t - b_t \parallel_2 \tag{4}$$

Vertex loss. When blendshapes loss is used alone, the neural network cannot learn the mapping relationship between the audio and the corresponding blendshapes parameters well, resulting in facial animation being unable to move. Please see Ablation experiment for details. By combining this loss, it can help the network learn the blendshapes parameters better. The Mean Squared Error (MSE) between the predicted vertex coordinates of each frame $\hat{v}_t$ and the ground truth v_t is vertex loss. The vertex loss can be expressed as:

$$L_{vertex} = \sum_{t=1}^{T} \parallel \hat{v}_t - v_t \parallel_2 \tag{5}$$

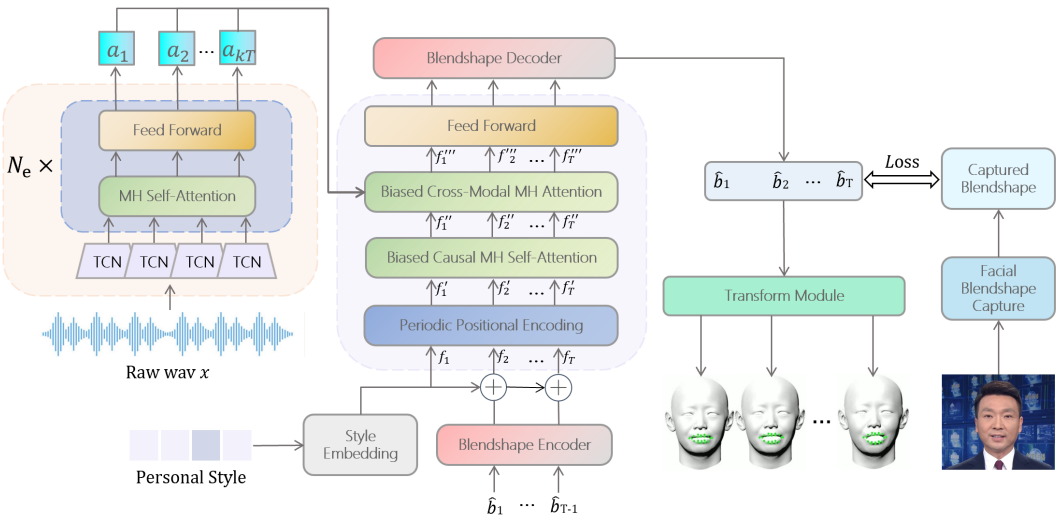

Figure 1: Overall network architecture. An encoder-decoder model with Transformer architecture takes raw audio as input and autoregressively generates a sequence of blendshapes parameters. Then convert it into 3D face meshes through the Transform module.

Experiments

Experimental Settings

Dataset. We use self-collected 3D dataset, CCTV News dataset for training and testing. The dataset provide the audio-blendshapes pairs of Chinese spoken utterances. This dataset is composed of 499 facial motion sequences from 8 subjects. Each sequence is captured at 25fps. Each 3D face mesh has 26,278 vertices and the corresponding 52 blendshapes parameters. Due to hardware limitations, we only used 249 audios during training. In order to verify the generalization ability of the model, we also collected 12 additional audio data read by ourselves. The calculation of errors in subsequent experiments will use these 12 audio data.

Baseline method. We compare our model with the state-of-the-art method, FaceFormer (Fan et al. 2022b), on self-collected 3D dataset.

Training details. During the training process, the model is optimized end-to-end using the Adam optimizer(Kingma and Ba 2014). The learning rate and batch size are set to 1E-4 and 1, respectively. The model is trained on a single NVIDIA RTX 2080 Ti, and the entire network takes approximately 14 hours (100 epochs) to train.

Quantitative evaluation

To measure lip synchronization, we calculated the lip vertex error (LVE). This evaluation metric computes the average L2 error of the lips in the test set. For a single frame, LVE is defined as the maximum L2 error among all lip vertices. We trained FaceFormer and our method on the self-collected 3D dataset. The blendshape parameters were converted into mesh vertices (26,278*3) corresponding to the Facescape model, which was used as ground truth. Tab. 1 shows LVE evaluation results. Faceformer has slight advantages over our approach. This is to be expected, because our method is to predict the blendshapes parameters and then fit the vertex coordinates, while faceformer directly predicts the vertex coordinates to be more accurate.

Table 1: Quantitative evaluation results of lip vertex error.

Method	**Lip Vertex Error**(mm)↓
Faceformer	**3.43465**
Ours	3.72332

Robustness analysis. When using the lip maximum L2 error metric, there may be a potential impact of outliers present in the dataset. To mitigate the impact of outliers and present a more comprehensive evaluation, we additionally computed the average lip vertex error (ALVE) for proposed method. For a single frame, ALVE is defined as the average L2 error among all lip vertices. In Tab. 2, we present the results of the ALVE obtained from our method and Faceformer.

Table 2: Quantitative evaluation results of average lip vertex error.

Method	**Average Lip Vertex Error**(mm)↓
Faceformer	**1.95755**
Ours	2.14514

Although our method is slightly inferior in error metrics, it has huge advantages in terms of the number of parameters of the model and the speed of inference. As shown in Table 3, our model size is 361MB and the number of parameters is 90,482,996. However, the model size of Faceformer is

Figure 2: Qualitative comparison of facial movement.

438MB and the number of parameters is 110,729,970. Judging from the number of parameters, it has been reduced by about 20%. Our model processes approximately 19.28 seconds of audio data per second in forward inference. Faceformer can only process about 9.16 seconds of audio data per second. Our inference speed is approximately twice that of Faceformer.

Table 3: Comparison of model size, number of parameters, and inference speed. Inference speed refers to how many seconds of audio data the model can process per second.

Method	Model size(MB)	Parameters	Speed
Faceformer	438	110,729,970	9.16
Ours	**361**	**90,482,996**	**19.28**

Qualitative evaluation

As audio and facial movements cannot be evaluated solely based on indicators and require human perceptual evaluation, we qualitatively evaluated our model from the perspective of lip synchronization. We compare our model with FaceFormer by feeding them the same audio input and generating corresponding facial animations. The results showed that the proposed model exhibited more pronounced lip movements and better alignment with human speech patterns. Figure 2 shows facial animation sequences generated by different methods for a certain Chinese word. When speaking the Chinese character ”hui guo”, our method generates facial movements with the lips opening wider. When speaking the Chinese character ’ma jiang pai’, our method generates better lip opening and closing effects.

Ablation experiment

When the loss function only uses blendshapes loss, the network cannot predict the corresponding blendshapes parameters well based on the audio. After two or three epochs of network training, the loss almost stopped declining and fell into a local minimum. The final generated facial animation is that the mouth is always closed.

In order to reduce blendshapes loss, vertex loss is introduced during training. After the network predicts the blendshapes parameters, it converts them into vertex coordinates through the transform module and then calculates vertex loss. With the guidance of vertex loss, blendshapes loss can be reduced well. Table 4 shows the error data under different loss strategies. LBE (lip blendshape error) represents the error of the correlation parameters with lip movement. FBE (full-face blendshape error) represents the full-face blendshape parameters error, that is, the error of all parameters. The combination of blendshape loss and vertex loss is smaller on LVE, but larger on LBE and FBE. But in fact, the animated mouth generated using only blendshape loss is always closed. Although its blendshape parameters is closer to the ground truth, the final effect is not good. Because the generated face is the complex weighted result of all parameters. Therefore, the quality of the method can only be judged through the LVE indicator and the effect of the video.

Table 4: LVE, LBE and FBE under different losses.

Method	LVE(mm)↓	LBE	FBE
blendshape only	4.03206	0.03798	0.07514
blendshape+vertex	**3.72332**	0.05069	0.07743

Conclusion

This paper proposes a method based on blendshapes parameters to generate speech-driven 3D facial animation. In the network, we introduce the transform module to realize the conversion of blendshapes parameters and vertex coordinates. Quantitative experimental results show that although our method has a relatively high error, our model has fewer parameters and faster inference speed. Qualitative experimental results show that our method can generate facial animations with lip movements more synchronized with audio.

References

Alghamdi, M. M.; Wang, H.; Bulpitt, A. J.; and Hogg, D. C. 2022. Talking Head from Speech Audio using a Pre-trained Image Generator. In *Proceedings of the 30th ACM International Conference on Multimedia*. ACM.

Baevski, A.; Zhou, H.; Mohamed, A.; and Auli, M. 2020a. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv:2006.11477.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020b. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Chen, L.; Cui, G.; Liu, C.; Li, Z.; Kou, Z.; Xu, Y.; and Xu, C. 2020. Talking-head Generation with Rhythmic Head Motion. arXiv:2007.08547.

Chen, L.; Li, Z.; Maddox, R. K.; Duan, Z.; and Xu, C. 2018. Lip Movements Generation at a Glance. arXiv:1803.10404.

Cudeiro, D.; Bolkart, T.; Laidlaw, C.; Ranjan, A.; and Black, M. J. 2019. Capture, Learning, and Synthesis of 3D Speaking Styles. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10093–10103.

Edwards, P.; Landreth, C.; Fiume, E.; and Singh, K. 2016. JALI: an animator-centric viseme model for expressive lip synchronization. *ACM Trans. Graph.*, 35: 127:1–127:11.

Fan, Y.; Lin, Z.; Saito, J.; Wang, W.; and Komura, T. 2022a. FaceFormer: Speech-Driven 3D Facial Animation with Transformers. arXiv:2112.05329.

Fan, Y.; Lin, Z.; Saito, J.; Wang, W.; and Komura, T. 2022b. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18770–18780.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, H.; Weise, T.; and Pauly, M. 2010. Example-Based Facial Rigging. *ACM Transactions on Graphics*, 29.

Massaro, D.; Cohen, M.; Tabain, M.; Beskow, J.; and Clark, R. 2001. Animated speech: Research progress and applications. *Audiovisual Speech Processing*.

Richard, A.; Zollhoefer, M.; Wen, Y.; de la Torre, F.; and Sheikh, Y. 2022. MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement. arXiv:2104.08223.

Taylor, S.; Kim, T.; Yue, Y.; Mahler, M.; Krahe, J.; Rodriguez, A. G.; Hodgins, J.; and Matthews, I. 2017a. A Deep Learning Approach for Generalized Speech Animation. *ACM Trans. Graph.*, 36(4).

Taylor, S. L.; Kim, T.; Yue, Y.; Mahler, M.; Krahe, J.; Rodriguez, A. G.; Hodgins, J. K.; and Matthews, I. 2017b. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36: 1 – 11.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2023. Attention Is All You Need. arXiv:1706.03762.

Xu, Y.; Feng, A. W.; Marsella, S.; and Shapiro, A. 2013a. A Practical and Configurable Lip Sync Method for Games. In *Proceedings of Motion on Games*, MIG '13, 131–140. New York, NY, USA: Association for Computing Machinery. ISBN 9781450325462.

Xu, Y.; Feng, A. W.; Marsella, S.; and Shapiro, A. 2013b. A Practical and Configurable Lip Sync Method for Games. In *Proceedings of Motion on Games*, MIG '13, 131–140. New York, NY, USA: Association for Computing Machinery. ISBN 9781450325462.

Yang, H.; Zhu, H.; Wang, Y.; Huang, M.; Shen, Q.; Yang, R.; and Cao, X. 2020a. FaceScape: a Large-scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction. arXiv:2003.13989.

Yang, H.; Zhu, H.; Wang, Y.; Huang, M.; Shen, Q.; Yang, R.; and Cao, X. 2020b. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 601–610.

Zhou, Y.; Xu, Z.; Landreth, C.; Kalogerakis, E.; Maji, S.; and Singh, K. 2018. VisemeNet: Audio-Driven Animator-Centric Speech Animation. arXiv:1805.09488.