Probabilistic Speech-Driven 3D Facial Motion Synthesis: New Benchmarks, Methods, and Applications

Karren D. Yang, Anurag Ranjan, Jen-Hao Rick Chang, Raviteja Vemulapalli, Oncel Tuzel
Apple
{karren_yang, anuragr, jenhao_chang, r_vemulapalli, ctuzel}@apple.com
https://github.com/apple/ml-audio2mesh

Abstract

We consider the task of animating 3D facial geometry from speech signal. Existing works are primarily deterministic, focusing on learning a one-to-one mapping from speech signal to 3D face meshes on small datasets with limited speakers. While these models can achieve high-quality lip articulation for speakers in the training set, they are unable to capture the full and diverse distribution of 3D facial motions that accompany speech in the real world. Importantly, the relationship between speech and facial motion is one-to-many, containing both inter-speaker and intra-speaker variations and necessitating a probabilistic approach. In this paper, we identify and address key challenges that have so far limited the development of probabilistic models: lack of datasets and metrics that are suitable for training and evaluating them, as well as the difficulty of designing a model that generates diverse results while remaining faithful to a strong conditioning signal as speech. We first propose large-scale benchmark datasets and metrics suitable for probabilistic modeling. Then, we demonstrate a probabilistic model that achieves both diversity and fidelity to speech, outperforming other methods across the proposed benchmarks. Finally, we showcase useful applications of probabilistic models trained on these large-scale datasets: we can generate diverse speech-driven 3D facial motion that matches unseen speaker styles extracted from reference clips; and our synthetic meshes can be used to improve the performance of downstream audio-visual models.

1. Introduction

Recently, there has been significant research interest in animating 3D faces from speech signals [8, 9, 15, 32, 42] with potential applications across immersive interactions, content creation and synthetic data generation. Most existing works approach this problem by learning a *deterministic* mapping from speech to 3D face meshes in a data-driven manner [8, 9, 15, 42], leveraging advancements in deep learning. These methods are typically optimized on small datasets containing 10-20 speakers [8, 16] and can achieve high-quality lip reconstruction for the speakers in the training dataset [8, 15, 42]. However, these methods fall short of capturing the *one-to-many* relationship between speech and realistic facial motions.

Animating 3D faces from speech is a complex problem. For a given speech utterance, there exists a multi-modal distribution of plausible facial motions capturing large variations in speaking style across a population. Even for a single speaker, the conditional distribution of facial motions given speech is multi-modal, capturing intra-speaker variations such as emotions [9] and other paralingustic cues that give nuance to the meaning of the speech. Modeling this complex, one-to-many relationship between speech and 3D facial motion necessitates a *probabilistic* approach, since approximating a multi-modal distribution with a deterministic point estimate leads to predicting the mean [8, 15] or a single mode [42] of the conditional distribution.

1.1. Challenges

Datasets. Learning this multi-modal distribution poses new challenges for the field of speech-driven 3D facial animation. First is the limitation of existing datasets. Building a useful probabilistic model that captures the wide variety of speech and facial motions requires a large amount of data from many speakers. However, existing public datasets are small and contain utterances from few speakers [8, 16], thus offering limited opportunity for learning diverse 3D facial motions. While a large-scale dataset is used in MeshTalk [32], this dataset is proprietary and not available to the research community.

Metrics. The second challenge is the lack of proper evaluation metrics for probabilistic speech-driven facial motion synthesis. Existing works use lip vertex error as the primary metric for evaluating lip synchronization [8, 32]. While lip vertex error is a useful proxy for lip articulation quality, it presumes a one-to-one relationship between speech and lip motion and penalizes realistic variations from the conditional mean. Other metrics such as upper-face dynamics deviation (FDD) have been proposed to measure the variability of the upper face, but they still compare the generated 3D facial motion against an absolute ground truth [42]. There is a need for metrics that are more suitable for evaluating lip quality and diversity in a probabilistic setting.

Modeling. Third, while learning to model the full distribution paves the way for realistic facial motions, it also opens the door to generating samples that are out of sync or of lower fidelity [31]. Most existing probabilistic models in other domains do not consider this problem, as their conditioning signals have weaker correlation with the synthesized content. Therefore, there is a need for modeling techniques that can achieve diverse facial motions while maintaining fidelity to the driving speech signal. Ensuring speech synchronization is made more difficult when also considering the need for other conditioning inputs, namely speaking style. Most existing works do not consider these challenges or interactions as they use one-hot speaker encodings and are not intended to generalize to unseen speaking styles.

1.2. Contributions

In this work, we address these challenges with new large-scale datasets, metrics, and modeling techniques for probabilistic speech-driven 3D facial animation.

Datasets. We propose to benchmark speech-driven 3D facial animation on two large-scale paired audio-mesh datasets derived from the VoxCeleb2 [6] video dataset using state-of-the-art monocular face reconstruction methods [17, 18]. These audio-mesh datasets contain thousands of speakers and are orders of magnitude larger than current public benchmarks [8, 16].

Metrics. We introduce metrics that are suitable for evaluating probabilistic models. We propose to quantify how well probabilistic models generate samples close to the ground truth lip motion, allowing a more comprehensive picture of lip articulation quality that takes the diversity of probabilistic models into account. We also train audio-mesh synchronization models and speaker recognition models to measure other aspects of generative quality, such as synchronization, realism, and diversity.

Modeling. We demonstrate a two-stage probabilistic autoregressive model over residual vector-quantized codes that achieves diverse generation while maintaining robust synchronization with speech. We also introduce simple but effective sampling strategies for trading off diversity for better lip precision and speech synchronization.

Results. We benchmark prominent deterministic (VOCA [8], Faceformer [15], CodeTalker [42]) and non-deterministic methods (MeshTalk [32]) on the large-scale datasets using suitable metrics. Our approach outperforms these existing methods, demonstrating the potential of probabilistic modeling. In perceptual studies, our approach is rated as producing more realistic lip and upper face motion, as well as more capable of capturing inter-speaker diversity (*i.e.,* matching reference clips) compared to deterministic models. Synthetic lip meshes generated from our method can be used to train downstream audio-visual models. On the challenging task of noisy audio-visual speech recognition on LRS3 [1], we improve relative WER by 11.3% compared to a model that is trained on the ground truth corpus and 47.0% compared to meshes from a deterministic model.

2. Related Work

Speech-driven face animation is a highly active field with extensive literature. Early viseme-based methods map the phonetic components of speech to their visual counterparts. More recent works have been either video-based methods that produce outputs in the pixel space, or 3D animation methods that drive facial motion as represented by 3D facial landmarks or meshes. There is overlap between these groups, in that some photorealistic methods also produce intermediate outputs such as facial landmarks or meshes. Below, we draw the distinction depending on whether the techniques mainly focus on the 3D facial geometry or on the photo-realistic video quality.

Viseme-Based Methods. Early methods use linguistic observations [3,27,36,44] to map from phoneme to viseme sequences. Phonemes are derived directly from text [2,13,14] or from speech via acoustic models [39]. Viseme sequences are subsequently translated to animations by morphing templates [13, 14, 23, 24] or 3D rigged models as in JALI [12]. More recently, deep learning methods have been introduced to learn the mapping function from phonemes to visemes [35,48]. While viseme-based methods provide interpretable controls over lip motion, their expressive power is limited; for example, they cannot produce subtle facial gestures in other regions of the face.

Video-Based Methods. There is extensive literature on synthesizing photorealistic talking heads from speech inputs. Most of these works synthesize 2D talking head videos [4, 5, 19, 28, 37, 40, 41, 43, 46, 47] and cannot easily be extended to 3D. Some methods incorporate neural rendering pipelines to synthesize 3D talking heads that can be rendered from different camera angles [20, 21, 45]. In general, these methods focus on realistically generating the pixels of a video, rather than the 3D facial motions.

3D Animation Methods. Early models are speaker-specific and cannot be used in more generic settings [25]. Early multi-speaker methods produce low-dimensional features such as blendshape coefficients [11]. Recent methods focus on animating the entire face from speech by directly

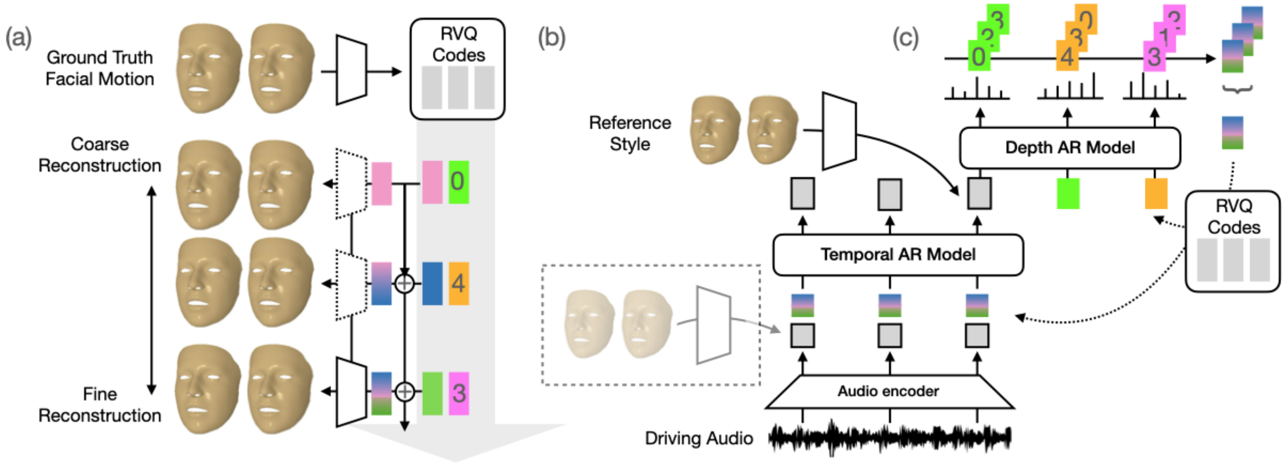


Figure 1. **Method Overview.** We learn a probabilistic model to synthesize 3D facial motion. (a) We first learn a residual vector-quantized codebook over the space of 3D facial motion. (b) We then train a two-stage, probabilistic auto-regressive model to predict these codes in a coarse to fine manner conditioned on audio and a reference speaker clip. (c) During inference, we propose sampling strategies to trade-off the diversity of the model in favor of improved lip fidelity. Colors - different tokens; color mixtures - token aggregation.

operating in the vertex space [8, 9, 15, 32, 42]. However, these methods mostly consider a deterministic formulation of the task. VOCA [8] and Faceformer [15] formulate speech-driven animation as a direct regression problem. In Meshtalk [32], lower face vertices are regressed from the speech signal through the bottleneck of a Gumbel-Softmax auto-encoder, and an autoregressive model is trained over the discrete codes. While the regression strengthens the correspondence between the speech signal and the generated lip motion, it limits the quality and diversity of the lower face. In CodeTalker [42], a discrete autoencoder is used to encode facial motion, and a separate model is used to regress the codes from audio. Different from these works, our model does not involve regressing facial motion from audio; rather, we model the full conditional distribution, which we show produces more diverse and realistic outputs.

In the context of dyadic 3D facial motion synthesis, Ng *et al.* [29] propose a probabilistic auto-regressive model for generating a listener's facial motions in a two-person conversation. However, the task differs from ours, in that while the listener's expressions are correlated with the speaker's voice and motions, this correlation is inherently weaker than in speech-driven facial motion.

Concurrent to our work, several groups have recently proposed probabilistic methods based on diffusion [33, 34], which is distinct from our auto-regressive approach. Comparison of diffusion and auto-regressive models in this context is interesting and should be explored in future work.

3. Approach

Our goal is to learn a probabilistic model $p_G(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{s})$ to synthesize 3D facial motion from speech, where $\boldsymbol{x} \in \mathbb{R}^{T\times 3V}$ is the target sequence of 3D mesh deformations, $\boldsymbol{y} \in \mathbb{R}^{T\times D_y}$ is the driving speech signal, and $\boldsymbol{s} \in \mathbb{R}^{T_s\times 3V}$ is a reference speaker sequence of 3D mesh deformations for controlling inter-speaker variation. We propose to first discretize the space of 3D facial motion using a *residual* vector-quantized (RVQ) codebook in a coarse-to-fine manner (Figure 1a, Section 3.1). Then, we propose an effective architecture for learning a two-stage probabilistic auto-regressive model over the codes (Figure 1b, Section 3.2). Finally, we propose sampling strategies to trade-off diversity for improved precision and speech synchronization, and propose a knowledge distillation strategy to amortize the sampling overhead (Section 3.3).

3.1. RVQ for 3D Facial Motion

Let $\mathcal{C}$ denote a fixed-size codebook with codes of size N_C. Residual vector quantization [26] is a discretization technique that recursively projects a vector $\mathbf{z} \in \mathbb{R}^{N_C}$ to the nearest code in $\mathcal{C}$ and takes the residual. After D steps, $\mathbf{z}$ can be represented by an ordered sequence of indices for the codes in $\mathcal{C}$, and the quantization of $\mathbf{z}$ up to depth d is represented by summing the codes corresponding to those indices. We apply RVQ to obtain a coarse-to-fine discretization of 3D facial motion by performing the above recursion within the latent space of a 3D facial motion autoencoder, as shown in Figure 1a. Specifically, we use an temporal convolutional encoder to map $\boldsymbol{x}$ to a latent embedding of motion, $Z \in \mathbb{R}^{T\times N_C}$. Each temporal index of Z is separately quantized using RVQ, and the quantized latent embedding of motion is decoded back to the 3D motion space using a convolutional decoder. The encoder and decoder of this autoencoder are jointly optimized via gradient updates to minimize reconstruction loss through the discrete code using a straight-through estimator [38]. The use of a commitment loss [26] to penalize the error of the quantization at every depth effectively ensures that the meshes can be reconstructed from the codes in a coarse-to-fine manner.

3.2. Two-Stage Probabilistic AR Model

From RVQ autoencoder, we obtain the codebook indices of a 3D mesh sequence $\boldsymbol{x}$. We denote these by a matrix $\mathbf{j}$, where j_{td} denotes the index for time point t and depth d. Next, we predict the individual code indices of $\mathbf{j}$ conditioned on $\boldsymbol{y}$ and $\boldsymbol{s}$,

$$\prod_{d=1}^{D}\prod_{t=1}^{T} p(j_{td}|\mathbf{j}_{<t}, \mathbf{j}_{t,<d}, \boldsymbol{y}_{\leq t}, \boldsymbol{s}), \tag{1}$$

using a two-stage [26] probabilistic Auto Regressive (AR) model consisting of a temporal model and a depth model (Figure 1b). The temporal model is an auto-regressive model that produces an audio-visual embedding for each time frame t capturing historical audio-visual context as well as the audio embedding from t:

$$\mathbf{h}_{av}[t] = \text{TemporalModel}(\boldsymbol{y}_{\leq t}, \tilde{e}(\mathbf{j}_{t-1}), \tilde{e}(\mathbf{j}_{t-2}), \cdots) \tag{2}$$

where $\tilde{e}(\mathbf{j}_t) := \sum_d e(j_{td})$ and $e(i)$ indicates the code in $\mathcal{C}$ corresponding to the i-th index. We experiment with both causal convolutional and transformer auto-regressive architectures for temporal model and find that the longer context of a transformer offers limited benefit when context information is provided through a reference style clip.

Subsequently, the depth model uses the audio-visual context captured in $\mathbf{h}_{av}[t]$ to generate each of the D code indices for the current time frame in an auto-regressive manner. The depth model consists of a masked self-attention transformer block which, at time frame t, operates along a length $D+1$ sequence v_t defined as: $v_{t1} = p_1 + E_s(\boldsymbol{s})$, $v_{t2} = p_2 + \mathbf{h}_{av}[t]$, and $v_{td} = p_d + \sum_{d'=1}^{d-1} e(j_{td'})$ for $d \geq 3$, where p_i denotes a learned positional encoding. The output of the depth model is a prediction of the conditional distribution of the next token.

$$p(j_{td}|j_{t,<d}, \mathbf{h}_{av}[t], \boldsymbol{s}) = \text{DepthModel}(v_{t,\leq d+1}) \tag{3}$$

Notice that we incorporate the encoded $\boldsymbol{s}$ as the first token input into the depth transformer, effectively shifting the standard input sequence by one. We find that incorporating speaker information as an input to the second-stage model, rather than as an input to in the first-stage model, which is more standard [26] and is showed as the grayed out box in Figure 1(b), is crucial for proper speech synchronization. As we show in Table 3, incorporating the speaker information into the first stage model rather than the second results in a decrease in synchronization. The two-stage auto-regressive model is trained end-to-end to minimize the cross-entropy loss, $-\mathbb{E}_{td} \log p(j_{td}|\mathbf{j}_{<t}, \mathbf{j}_{t,<d}, \boldsymbol{y}_{\leq t}, \boldsymbol{s})$, in a teacher-forcing manner.

Discussion. Our motivation for using RVQ over other types of discretization include greater representational power as well as the ability to reduce the number of codes predicted during inference for greater speed, since the RVQ codes are ordered from coarse-to-fine (see Supplement for details).

3.3. Trading off Diversity

During inference, we can sample from the conditional distribution of facial motions as shown in Equation 1. This achieves good results, but we also want to control the diversity/variability of the synthesis. In particular, the training loss forces the probabilistic AR to capture the entire training distribution of codes, which is noisy and can result in sampling codes that are less faithful to the conditioning speech during inference. Therefore, we provide some sampling strategies to trade-off diversity for fidelity to the speech signal: (1) KNN-based sampling, (2) code averaging, and (3) rejection sampling using a pre-trained synchronization network. As shown in Figure 1c, we sample multiple codes and aggregate their embeddings before passing the result as the next input to the temporal model.

KNN-based sampling. For simplicity of notation, let $e_t := \tilde{e}(\mathbf{j}_t)$ denote the sampled and reconstructed quantized embedding for time t. We replace the sampled code at time step t with the mean of a local Gaussian approximated from its nearest neighbors on the sampling manifold. Let $\mathcal{E}$ denote a set of N codes sampled at time step t. We take the estimate $\hat{e}_t$ to be the mean of the set $\{e \in \mathcal{E} \mid |e - e_t| \leq |\text{KNN}_k(e_t, \mathcal{E}) - e_t|\}$, where $\text{KNN}_k(e_t, \mathcal{E})$ denotes the k-th nearest neighbors of e_t in $\mathcal{E}$. The replacement code $\hat{e}_t$ is projected to the discrete codebook.

Code averaging. We replace the sampled code e_t with a embedding $\hat{e}_t$ given by the mean of $\mathcal{E}$, a set of N codes sampled at time step t. The averaged embedding $\hat{e}_t$ is projected to the discrete codebook.

SyncNet-based Sampling. Inspired by classifier-based rejection sampling in image synthesis, we propose a simple sampling scheme based on a pretrained synchronization network. Specifically, at each time point t, we sample and decode a set of N codes. Each code e_t is decoded by the RVQ autoencoder and scored using a pretrained synchronization network.

While these strategies increase the computational overhead of inference, we can amortize them by distilling the modified sampling distributions into a student network that can be run with no additional cost during inference. We do so by relabeling the code inputs as well as targets of the depth network by the ones obtained from discretizing the aggregated samples (see Supplemental Materials for details).

Dataset	# Mesh Sequences	# Speakers
VOCASet	480	12
BIWI	1109	14
VoxCeleb2 (Mesh)	>1M	6,112

Table 1. **Comparsion of Different Benchmark Datasets for Speech-Driven 3D Facial Animation.** Our proposed benchmark datasets of meshes reconstructed from VoxCeleb2 are significantly larger than existing benchmark datasets.

4. Experiments

4.1. Benchmark Datasets

Most of the existing works on speech-driven 3D facial motion synthesis use VOCASet [8] and BIWI [16] for benchmarking. These datasets are small with a limited number of speakers, and models are often trained and evaluated in a speaker-specific manner on these datasets. Because of their small scale and limited speaker diversity, these datasets do not fully capture the complex relationship between speech and facial motions. While MeshTalk [32] uses a large, multi-speaker dataset to train their model, their dataset is proprietary and not available for public use.

To address this issue, we introduce two large-scale audio-mesh benchmark datasets. These datasets are created by processing videos from the publicly-available VoxCeleb2 video dataset [6] using two monocular face reconstruction methods: DECA [17], a state-of-the-art method for face reconstruction, and SPECTRE [18], a recent method that holds the state-of-the-art for preserving visual speech information. These two datasets contain face meshes at different granularity enabling us to assess how well different speech-driven facial motion synthesis methods fare on different types of meshes. Table 1 shows the statistics of the different datasets. Note that VoxCeleb2 is orders of magnitude larger than the existing benchmark datasets, enabling the development of models that capture speaker diversity reflective of a real-world population.

4.2. Metrics

Lip Vertex Error. In existing works, lip vertex error is used as the main proxy for lip articulation quality. This metric is calculated as

$$\ell_{vertex}(x, \hat{x}) := \max_{t,i\in\text{lip}} ||\boldsymbol{x}_{ti} - \hat{\boldsymbol{x}}_{ti}||_2 \qquad (4)$$

where x is the ground truth mesh, $\hat{x}$ is the synthesized mesh, the maximum is taken over all lip vertices and time frames for a given mesh sequence. However, there is a distribution of possible lip vertex positions for a given individual and utterance, and the ground truth is only one sample from this distribution. Lip vertex error does not reflect that a probabilistic model may correctly capture multiple modes that include the ground truth, but receive a large lip vertex error by sampling a different mode. While the lip vertex error measures the precision of the model, or how close every sample is to the ground truth, a more suitable metric for a probabilistic model may be whether any one of several samples, or their mean, is close to the ground truth, which mitigates the impact of diversity on this metric.

Coverage Error. To provide a notion of how close the ground truth is to the sampling distribution of a probabilistic model, we propose to generate a set of samples $\mathcal{S}$ and computing the closest distance to the ground truth:

$$\ell_{cover} := \min_{\hat{x}\in\mathcal{S}} \ell_{vertex}(x, \hat{x}).$$

Intuitively, a probabilistic model with small ℓ_{cover} has a mode that is close to the ground truth, even if a generated sample is not.

Mean Estimate Error. Finally, we also propose to compute the lip vertex error over the mean of $\mathcal{S}$, *i.e.*, $\ell_{mean} := \ell_{vertex}(x, \mathbb{E}_{\hat{\mathcal{S}}}\hat{x})$, to assess how close the mean of the sampling distribution is to the ground truth. Both coverage error and the error of the mean error better reflect whether a probabilistic model is capable of generating the ground truth lip sequence better than computing error from one random sample. Note that all three lip errors are the same for deterministic methods, as they are only capable of generating the same sample.

SyncNet Score. While lip vertex errors measure how close generated lip articulations are to the ground truth, they do not reflect whether a particular 3D mesh sequence falls into the possible distribution of facial motions conditioned on a speech utterance. We propose to learn this distribution by training an speech-mesh synchronization network that scores how well a mesh corresponds to a given audio, analogous to the lip synchronization metric used in speech-driven video synthesis [7]. Specifically, we pretrain two different synchronization networks to assess the alignment between a mesh sequence and an audio signal. In the first network, a multimodal fusion network is used to merge mesh and audio embeddings along the temporal dimension, and a score is computed from the merged embeddings using a linear layer. In the second network, the score is computed directly through the cosine similarity of the normalized mesh and audio embeddings. Both networks are optimized using an InfoNCE contrastive loss [30], and perform well at detecting temporal as well as semantic alignment between audio and 3D face meshes (see Supplemental Materials).

SyncNet Frechet Distance (SyncNet-FD). Beyond measuring the quality of speech synchronization, we also want to measure how well the speech-related facial motions generated by a model capture the realism and diversity of the real distribution of such motions. To do so, we compute the Frechet distance between 1000 SyncNet embeddings of real and generated mesh sequences from our two pretrained speech-mesh synchronization networks.

Style Cosine Similarity and Rank. While the above metrics provide different measures of how well models generate 3D facial motions corresponding to speech, we also want to measure how well models are able to replicate the diverse speaking styles within the datasets. To do so, we train a speaking style recognition model based on ArcFace [10], using 3D facial motions as input (*i.e.,* the deformation of ground truth meshes from the neutral templates). We evaluate how well the models are able to replicate a specific individual's speaking style by computing the cosine similarity between the embeddings of the reference speaker mesh sequences and the generated mesh sequences. We also compute the rank of the similarity relative to the similarity of all the other speakers in the training set. Details of the implementation and performance of the recognition model are provided in the Supplemental Materials.

Style Frechet Distance (Style-FD). Finally, to assess the diversity of speaking styles produces by the model and how well the distribution matches the speaking styles of the real data, we compute the Frechet Distance between the recognition model embeddings of the real and generated mesh sequences.

4.3. Quantitative Results

Figure 2 shows the results of our comprehensive benchmark. We train recent deterministic and probabilistic methods on our DECA and SPECTRE meshes and evaluate them using our proposed metrics. Overall, our method outperforms the existing methods on realism/diversity (as measured by averaged FD score), speech synchronization, and lip coverage and mean estimate errors. We provide a thorough discussion below.

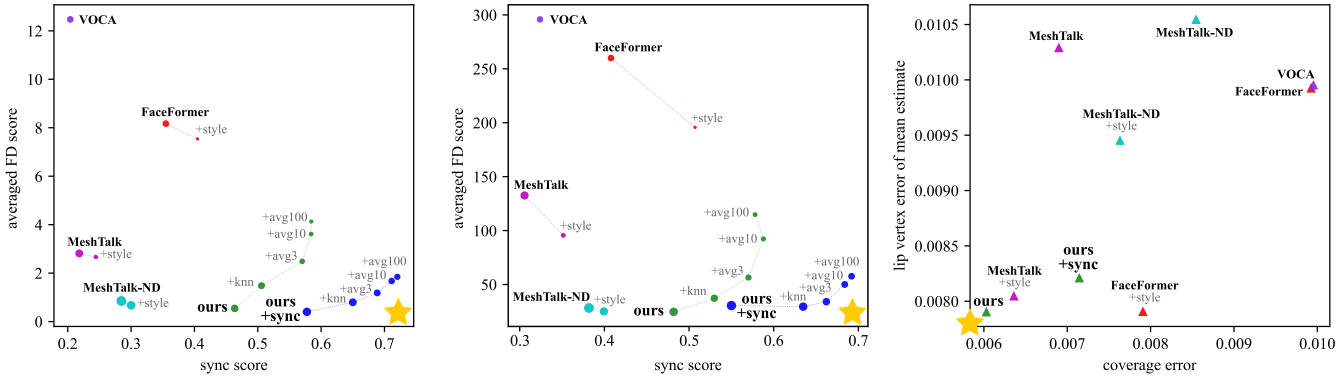
Ours *vs*. Deterministic Methods. Existing deterministic methods[1] suffer in realism/diversity as measured by averaged FD (y-axis, lower is better) on both DECA (Figure 2a) and SPECTRE meshes (Figure 2b). Specifically, VOCA [8] and Faceformer [15] are deterministic methods that directly regress 3D mesh vertices on speech, either using a sliding window (VOCA) or an auto-regressive transformer (FaceFormer). Both methods are susceptible to the over-averaging effect of the regression loss, which is exacerbated by training on large-scale datasets. We observe that FaceFormer produces less stiff motions compared to VOCA, due to conditioning on a longer context provided by auto-regressive modeling, resulting in higher synchronization scores. We add conditioning on a reference speaker sequence to FaceFormer to further reduce the distribution of possible facial motions (FaceFormer+Style). This improves its scores across all metrics, particularly on the SPECTRE meshes that are more detailed, but does not resolve the realism/diversity gap.

Our method also outperforms VOCA and Faceformer on speech synchronization, as measured by the sync score in Figures 2(a-b) (x-axis, higher is better). On the SPECTRE meshes (b), FaceFormer+Style achieves higher SyncNet score compared to default sampling from our probabilistic model, but we can achieve better results using our proposed sampling strategies, as illustrated by the green and blue markers (see later section for discussion). For lip vertex errors, VOCA and FaceFormer both achieve lower ℓ_{vertex} (Figure 2(a-b), marker size, smaller is better), but this is mainly because this metric penalizes the diversity of samples generated by our probabilistic modeling. When we compute the lip vertex error over the average of many samples from our model (Figure 2(c), y-axis, lower is better) (ℓ_{mean}, $|\mathcal{S}| = 100$), effectively reducing the effect of sampling diversity, we outperform VOCA and FaceFormer and are able to match the lip vertex error of FaceFormer+Style. Furthermore, our model achieves better coverage error than FaceFormer+Style (Figure 2c, x-axis, lower is better), indicating that our sampling distribution is actually much closer to the ground truth lip vertices.

Ours *vs*. MeshTalk. MeshTalk [32] is a two-stage method that first learns a discrete Gumbel-Softmax autoencoder [22] that disentangles upper and lower face motion, then trains a probabilistic auto-regressive model over the discrete codes using a convolutional architecture. While the second stage model is probabilistic, disentangling the lower face involves regressing the vertices from audio over sliding windows, similar to VOCA. We observe that MeshTalk is susceptible to the same over-smoothing effects on the lower face, achieving similar synchronization scores to VOCA in Figure 2(a-b). Overall, our method achieves better synchronization as well as realism/diversity compared to MeshTalk and MeshTalk+Style, as reflected in higher sync scores and lower averaged FD score. For lip vertex error (ℓ_{vertex}), our meshes are more diverse, and thus deviate from the ground truth meshes more than MeshTalk+Style. However, we achieve better coverage error as well as mean estimate error, suggesting that while our results are more diverse, our sampling distribution is actually closer to the ground truth.

We also train a version of MeshTalk without the regression loss in the codebook (MeshTalk-ND, Meshtalk-ND+Style), for a more direct comparison to another probabilistic auto-regressive model that predicts discrete latent codes. Compared to the original version, MeshTalk-ND and MeshTalk-ND+Style are more diverse, as evidenced by lower SyncNet-FD scores, and they are not susceptible to smoothing of the lower face, as evidenced by improved synchronization scores. However, the quality of the lip articu-

[1] We defer discussion of CodeTalker [42] to the Supplement.



(a) on DECA meshes

(b) on SPECTRE meshes

(c) coverage and mean estimate error

Figure 2. **Benchmark results.** We evaluate all methods on the aggregate SyncNet score and averaged FD score on (a) DECA and (b) SPECTRE meshes from VoxCeleb2. The size of the dots indicates the lip vertex error. Averaged FD score refers to the average between both SyncNet-FD scores. (c) shows the coverage error and the lip vertex error of the mean estimate over 100 samples per speech input on DECA meshes from VoxCeleb2. The yellow star indicates the direction of the best methods. For averaged FD score, lower is better. For sync score, higher is better. For both coverage and mean estimate error, lower is better. See supplementary material for the complete table.

lation suffers. Note that MeshTalk-ND+Style cannot cover the ground truth lip sequences as well as MeshTalk+Style or our approach, even though ours is just as diverse. Our approach also achieves higher synchronization scores. This demonstrates the effectiveness of our probabilistic model design choices in maintaining faithfulness to the driving speech signal.

Ours *vs*. Diffusion Methods. Concurrent to our work, several groups have introduced probabilistic methods based on diffusion [33, 34]. We performed preliminary comparisons with diffusion models by training FaceDiffuser [33] on our data. Qualitatively, we observed that FaceDiffuser captures more diversity than deterministic methods like Faceformer, but less than our approach. This may be due to limited representational capacity of the denoising function, which was designed for smaller datasets [33].

Trading off Diversity for Fidelity. By design, our probabilistic model learns the entire training distribution of RVQ codes, which is noisy and can result in sampling codes that are less faithful to the conditioning speech during inference. The results in Figure 2(a-b) show that we are able to trade-off diversity for greater fidelity using the strategies in Section 3.3 with (blue) or without (green) SyncNet-based rejection sampling. KNN-based sampling achieves a mild trade-off, as code aggregation is based on a local Gaussian approximation. Code averaging achieves a larger trade-off, as the model samples codes that are closer to the conditional mean $\mathbb{E}[\boldsymbol{x}_t|\boldsymbol{x}_{<t}, \boldsymbol{y}, \boldsymbol{s}]$. When averaging between large numbers of codes, eventually the synchronization score decreases due to over-smoothing.

DECA	**Style Cosine Similarity ↑**	**Style Rank ↓**	**Style FD↓**
FaceFormer+Style	0.127	1596.422	58.652
MeshTalk+Style	0.229	1135.0	38.535
MeshTalk-ND+Style	0.629	53.8	**17.068**
Ours	**0.707**	**7.3**	21.038
GT	0.7644	10.691	-
SPECTRE	**Style Cosine Similarity↑**	**Style Rank ↓**	**Style FD↓**
FaceFormer+Style	0.237	650.128	41.062
MeshTalk+Style	0.231	955.3	69.224
MeshTalk-ND+Style	0.609	38.8	**20.560**
Ours	**0.673**	**20.5**	23.533
GT	0.7522	4.982	-

Table 2. **Style Similarity Scores** show that our probabilistic approach can synthesize facial motion closer to the reference style compared to other deterministic methods. See text for details.

Speaker Style Evaluation. Next, we evaluate the ability of our method to generate the diverse speaking styles of unseen speakers, provided with a reference clip from the target speaker. The results are shown in Table 2, and we compare to other methods that are also trained using a reference clip. Overall, we find that FaceFormer+Style and MeshTalk+Style, which both employ some form of regression from speech in the training stage, are unable to match the speaking style of the target speakers due to over-smoothing and loss of diversity in the facial motions. This is reflected not only in the style cosine similarity, but also in the higher Style-FD. As previous works have noted that recognition networks may be sensitive to slight perturbations introduced by discrete coding schemes [31], we evaluate our method and MeshTalk-ND+Style on the decompressed ground truth meshes of their respective codebook. This improves the style matching scores of both models, which approach the scores of the real ground truth meshes.

Key Design Choices for AR Modeling. One challenge of our task is capturing the diverse facial motions correspond-

Method	**Style Cosine Similarity** ↑	**Sync Score** ↑	**Sync FD** ↓
AR-ConvNet (no style)	-	0.217	9.58
AR-Transformer (no style)	-	0.442	4.21
AR-Transformer+ES	0.315	0.287	2.95
Ours	0.298	0.4634	3.21

Table 3. **Key Ablations** of our model. We show that the design of the auto-regressive model is crucial for proper synchronization.

	Style Matching	**Lip Realism**	**Upper Face Realism**
Ours *vs.* VOCA	85.1/8.5/6.4	70.2/17.0/12.8	80.9/4.3/14.9
Ours *vs.* FaceFormer	74.4/20.5/5.1	59.0/35.9/5.1	71.1/26.3/2.6
Ours *vs.* CodeTalker	78.8/9.1/12.1	87.9/3.0/9.1	90.9/3.0/6.1
Ours *vs.* Faceformer+Style	75.0/11.1/13.9	86.1/8.3/5.6	94.4/2.8/2.8

Table 4. **Results of a Perceptual Study**. Results show percentage of survey respondents who preferred Ours / Baseline / Neither on each of the categories. For style matching, users were provided a reference clip in addition to two videos and asked which one matched the style in the clip better, which one had more realistic lower lip motion, and which one had more realistic upper face motion.

Training Data Type	Training Data Corpus	WER ↓
Audio-only	LRS3 trainval+pretrain	18.7
Real AV	LRS3 trainval	30.7
+ Faceformer Synthetic Dataset	LRS3 pretrain	13.4
+ Ours Synthetic Dataset	LRS3 pretrain	**7.1**
+ Real AV	LRS3 pretrain	8.0

Table 5. **Synthetic Data Generation for AVSR** Training an audio-visual speech recognition model on synthetic meshes generated by our model improves WER over training on meshes extracted from ground truth videos.

ing to speech while maintaining faithfulness to speech signal. In Table 3, we show the results of ablation studies that highlight our key design choices. First, using a convolutional architecture for the auto-regressive modeling, as in MeshTalk, results in significantly worse sync scores. Second, incorporating style information early in the temporal AR model, rather than the depth AR model, as done in many works that condition on global embeddings, significantly impairs the synchronization score.

4.4. Applications

We showcase two useful applications of a probabilistic model trained on a diverse large-scale dataset. The first application is the ability to generate more natural and realistic 3D facial motions that capture a diversity of real-world speaking styles, including being able to match the style from a reference clip. We show the results of user ratings in Table 4, illustrating that our approach is strongly preferred over prominent deterministic methods trained on smaller, high-quality datasets, as well as FaceFormer+Style trained on our large-scale datasets. A separate study on lip synchronization can be found in the Supplemental Materials. Second, we demonstrate the utility of probabilistic methods for generating synthetic training data for downstream audio-visual tasks. Specifically, we consider the challenging task of noisy audio-visual speech recognition (noisy-AVSR) on the Lip Reading Sentences 3 (LRS3) dataset. High-quality synthetic training data is immensely useful for audio-visual speech recognition, not only because labeled audio-visual corpora are limited, but also because there may be privacy concerns with training and deploying a model on real user data. We show that synthetic data from our speech-driven 3D facial animation model can greatly improve the performance of such audio-visual models, even compared to training on the ground truth visual data. We use our model trained on SPECTRE meshes to generate a large, synthetic 3D facial mesh dataset corresponding to the audio in the “pretrain” subset of the LRS3 dataset and use the detailed lip meshes as input to the downstream model. As shown in Table 5, training an audio-visual speech recognition model on this synthetic visual corpus improves relatively the WER of the model on the test set of LRS3 by 11.3% compared to training on the ground truth lip meshes, and by 47% compared to training on meshes generated by FaceFormer (also trained on SPECTRE meshes). Beyond creative applications, this demonstrates the practical usage of non-deterministic 3D facial mesh synthesis methods for training downstream audio-visual models.

5. Conclusion

In this work, we propose new large-scale benchmarks and methodology to address the task of probabilistic speech-driven 3D facial motion synthesis. We show the advantages of probabilistic approaches to this task in capturing diversity and propose a careful model design and sampling strategies to ensure strong lip synchrony. We benchmark existing deterministic methods on our large-scale dataset and show that our probabilistic approach outperforms them across metrics capturing realism, diversity and lip synchronization. Overall, our work provides useful large-scale benchmarks and metrics for other researchers working on this task.

Limitations. Our benchmark dataset relies on state-of-the-art monocular face reconstruction techniques [17, 18] and the VoxCeleb2 video dataset [6]. The quality of the ground truth face meshes is limited and noisier compared to those reconstructed from high-resolution multi-view videos. We leave the problem of reconstructing facial motion robustly from large-scale, in-the-wild training data to future work.

Acknowledgment. We thank Zak Aldeneh, Barry Theobald, Masha Fedzechkina Donaldson, Dianna Yee, and Hadi Pour Ansari for discussions and feedback on the paper.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. 2

[2] Robert Anderson, Bjorn Stenger, Vincent Wan, and Roberto Cipolla. Expressive visual text-to-speech using active appearance models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3382–3389, 2013. 2

[3] Yong Cao, Wen C Tien, Petros Faloutsos, and Frédéric Pighin. Expressive speech-driven facial animation. *ACM Transactions on Graphics (TOG)*, 24(4):1283–1302, 2005. 2

[4] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generationwith dynamic pixel-wise loss. *arXiv preprint arXiv:1905.03820*, 2019. 2

[5] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? *arXiv preprint arXiv:1705.02966*, 2017. 2

[6] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 2, 5, 8

[7] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 251–263. Springer, 2017. 5

[8] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019. 1, 2, 3, 5, 6

[9] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael J Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. *arXiv preprint arXiv:2306.08990*, 2023. 1, 3

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 6

[11] Zhigang Deng, Pei-Ying Chiang, Pamela Fox, and Ulrich Neumann. Animating blendshape faces by cross-mapping motion capture data. In *Proceedings of the 2006 symposium on Interactive 3D graphics and games*, pages 43–48, 2006. 2

[12] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on graphics (TOG)*, 35(4):1–11, 2016. 2

[13] Tony Ezzat and Tomaso Poggio. Miketalk: A talking facial display based on morphing visemes. In *Proceedings Computer Animation'98 (Cat. No. 98EX169)*, pages 96–102. IEEE, 1998. 2

[14] Tony Ezzat and Tomaso Poggio. Visual speech synthesis by morphing visemes. *International Journal of Computer Vision*, 38:45–57, 2000. 2

[15] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18780, 2022. 1, 2, 3, 6

[16] Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6):591–598, 2010. 1, 2, 5

[17] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 2, 5, 8

[18] Panagiotis P Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Spectre: Visual speech-informed perceptual 3d facial expression reconstruction from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5744–5754, 2023. 2, 5, 8

[19] Jiazhi Guan, Zhanwang Zhang, Hang Zhou, Tianshu Hu, Kaisiyuan Wang, Dongliang He, Haocheng Feng, Jingtuo Liu, Errui Ding, Ziwei Liu, et al. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1515, 2023. 2

[20] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021. 2

[21] Ricong Huang, Peiwen Lai, Yipeng Qin, and Guanbin Li. Parametric implicit face representation for audio-driven facial reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12759–12768, 2023. 2

[22] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 6

[23] Gregor A Kalberer, Pascal Müller, and Luc Van Gool. Speech animation using viseme space. In *VMV*, pages 463–470, 2002. 2

[24] Gregor A Kalberer and Luc Van Gool. Face animation based on observed 3d speech dynamics. In *Proceedings Computer Animation 2001. Fourteenth Conference on Computer Animation (Cat. No. 01TH8596)*, pages 20–251. IEEE, 2001. 2

[25] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 2

[26] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022. 3, 4

[27] DW Massaro, MM Cohen, M Tabain, J Beskow, and R Clark. Animated speech: research progress and applications. 2012. 2

[28] Gaurav Mittal and Baoyuan Wang. Animating face using disentangled audio representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3290–3298, 2020. 2

[29] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20395–20405, 2022. 3

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[31] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 2, 7

[32] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1173–1182, 2021. 1, 2, 3, 5, 6

[33] Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pages 1–11, 2023. 3, 7

[34] Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Gaetan Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong-jin Liu. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *arXiv preprint arXiv:2310.00434*, 2023. 3, 7

[35] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. A deep learning approach for generalized speech animation. *ACM Transactions On Graphics (TOG)*, 36(4):1–11, 2017. 2

[36] Sarah L Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. Dynamic units of visual speech. In *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, pages 275–284, 2012. 2

[37] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024. 2

[38] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 4

[39] Ashish Verma, Nitendra Rajput, and L Venkata Subramaniam. Using viseme based acoustic models for speech driven lip synthesis. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 5, pages V–720. IEEE, 2003. 2

[40] Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou Li. Seeing what you said: Talking face generation guided by a lip reading expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14653–14662, 2023. 2

[41] Jiayu Wang, Kang Zhao, Shiwei Zhang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Lipformer: High-fidelity and generalizable talking face generation with a pre-learned facial codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13844–13853, 2023. 2

[42] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023. 1, 2, 3, 6

[43] Chao Xu, Junwei Zhu, Jiangning Zhang, Yue Han, Wenqing Chu, Ying Tai, Chengjie Wang, Zhifeng Xie, and Yong Liu. High-fidelity generalized emotional talking face generation with multi-modal emotion space learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6609–6619, 2023. 2

[44] Yuyu Xu, Andrew W Feng, Stacy Marsella, and Ari Shapiro. A practical and configurable lip sync method for games. In *Proceedings of Motion on Games*, pages 131–140. 2013. 2

[45] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, Jinzheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*, 2023. 2

[46] Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3867–3876, 2021. 2

[47] Zhimeng Zhang, Zhipeng Hu, Wenjin Deng, Changjie Fan, Tangjie Lv, and Yu Ding. Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3543–3551, 2023. 2

[48] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics (TOG)*, 37(4):1–10, 2018. 2