Supervising 3D Talking Head Avatars with Analysis-by-Audio-Synthesis

Radek Daněček Carolin Schmitt Senya Polikovsky Michael J. Black
{radek.danecek, cschmitt, senya, black}@tuebingen.mpg.de
Max Planck Institute for Intelligent Systems, Tübingen, Germany

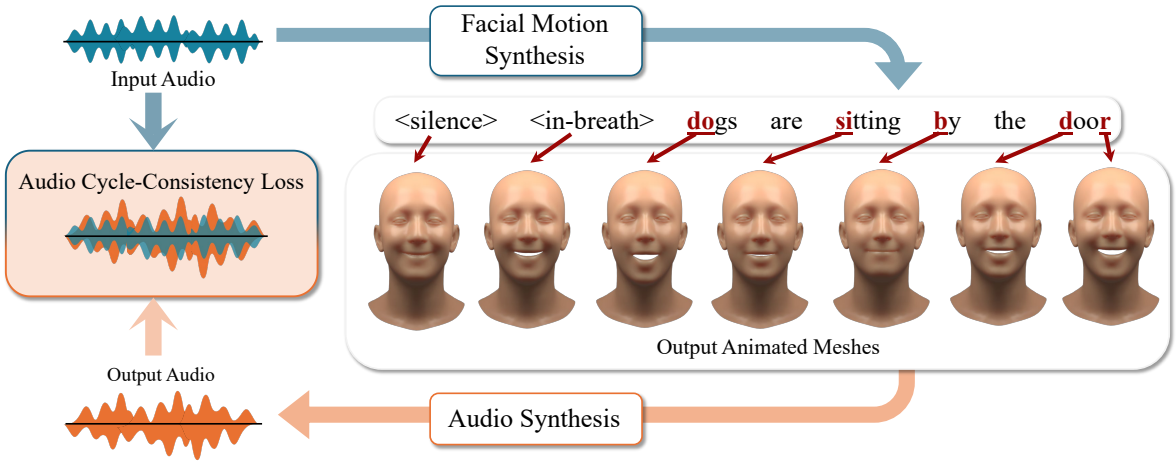

Figure 1. THUNDER introduces a new paradigm for stochastic generation of 3D talking head avatars from speech with accurate lip articulation and diverse facial expressions. Given an input audio, a 3D animation is generated. The animation is then fed into an audio synthesis (mesh-to-speech) model, which generates an output audio representation. The input and output audio representations are compared, creating a novel audio-consistency supervision loop, which we coin as *analysis-by-audio-synthesis*.

Abstract

In order to be widely applicable, speech-driven 3D head avatars must articulate their lips in accordance with speech, while also conveying the appropriate emotions with dynamically changing facial expressions. The key problem is that deterministic models produce high-quality lip-sync but without rich expressions, whereas stochastic models generate diverse expressions but with lower lip-sync quality. To get the best of both, we seek a stochastic model with accurate lip-sync. To that end, we develop a new approach based on the following observation: if a method generates realistic 3D lip motions, it should be possible to infer the spoken audio from the lip motion. The inferred speech should match the original input audio, and erroneous predictions create a novel supervision signal for training 3D talking head avatars with accurate lip-sync. To demonstrate this effect, we propose THUNDER (Talking Heads Under Neural Differentiable Elocution Reconstruction), a 3D talking head avatar framework that introduces a novel supervision mechanism via differentiable sound production. First, we train a novel mesh-to-speech model that regresses audio from facial animation. Then, we incorporate this model into a diffusion-based talking avatar framework. During training, the mesh-to-speech model takes the generated animation and produces a sound that is compared to the input speech, creating a differentiable analysis-by-audio-synthesis supervision loop. Our extensive qualitative and quantitative experiments demonstrate that THUNDER significantly improves the quality of the lip-sync of talking head avatars while still allowing for generation of diverse, high-quality, expressive facial animations.

1. Introduction

Animating 3D faces from speech has many applications. Examples include the entertainment industry (games and movies), applications in AR/VR/XR (virtual telepresence), e-commerce and perhaps future embodied digital assistants. In order to be widely applicable, the 3D head avatars must, however, appear and act believably and feel alive. While the methods to animate 3D avatars from speech have come a long way in recent years, there is still a considerable realism gap between real humans and 3D avatars.

Seminal approaches such as VOCA [16] or Karras et al. [45], and others [18, 26, 61] are deterministic regressors, that predict 3D vertices from the input audio. Despite achieving good lip-sync, these methods produce facial an-

imations without emotion [16, 26] or that are unnaturally static [18, 61]. This is in part due to the lack of large-scale and expression-rich training data and in part because a deterministic regressor cannot accurately produce animations of facial motions that only weakly correlate with the audio (such as eyebrow motion, eyeblinks, etc.). Hence, in order to generate animations that appear lively and believable, a good 3D speech-driven system must model the problem in a non-deterministic manner so that it can capture the many-to-many mapping between speech and facial animation.

Consequently, recent methods model the problem of 3D speech-driven animation stochastically using relatively small datasets of high-quality 3D scans [69, 73, 83, 91]; the small training size limits diversity and expressiveness. Other methods leverage lower-quality pseudo ground-truth (GT) extracted from larger-scale video datasets [74, 98]. In pursuing diversity, however, stochastic methods vary the lip motions in ways that may deviate from the audio signal.

The fundamental issue is that lip motions, for the same audio, vary with expression. We seek a method that generates diverse expressions but is faithful to the audio. So we ask the question: "What does it mean to be faithful to the audio?" This leads us to our key observation; that is, we seek 3D lip motions that *actually reproduce the input audio*. We formalize this notion with a **new form of supervision** for 3D speech-driven avatars using a novel **mesh-to-speech** model (M2S); see Fig. 1. Inspired by recent silent-video-to-speech (SV2S) methods [12, 47, 48], we propose, to our knowledge, the first M2S model. Our M2S model regresses the speech audio (or a representation thereof) solely from the animated 3D face. In other words, every generated animation now also produces a sound. Our rationale is simple: the more similar the produced sound is to the input speech, the more plausible the animation. Furthermore, failure to produce the correct sound should provide a good supervision signal.

Our system consists of two stages. In the first stage, we train the M2S model to regress audio from facial motion. In the second stage, we train a diffusion model to output 3D facial animation from speech. In this stage, we utilize the frozen M2S model. The generated facial animation is fed into the M2S model, effectively reproducing the spoken audio. This generated audio (or representation thereof) is then compared to the input audio, creating a self-supervised training loop; this is illustrated in Fig. 1. We refer to this as *analysis-by-audio-synthesis*. Our experiments show that leveraging the M2S model results in much better lip-sync quality, while still enabling stochastic production of expressive speech-driven animation.

In summary, our contributions are: (1) A novel **mesh-to-speech** model that regresses sound from 3D facial animations. (2) A new form of **cross-modal audio-based self-supervision** via comparison of the representations of the input speech and the animation-produced output speech. (3) **THUNDER**, a 3D speech-driven animation method based on diffusion, capable of generating a variety of facial expressions while maintaining accurate lip-sync. The code, data annotations and models will be made publicly available at https://thunder.is.tue.mpg.de/.

2. Related work

2.1. Speech-driven 3D animation

The field of speech-driven 3D facial animation has made significant progress in the last two decades [7, 13, 22, 23, 77, 92]. Here we focus on the recent deep-learning based line of work [2, 3, 16, 18, 24–26, 45, 57, 58, 61–63, 69, 73, 74, 78, 83, 91, 95, 98, 99].

Deterministic neural methods. Karras et al. [45] were the first to utilize deep learning by training a temporal convolutional network (TCN) to predict 3D face vertices. VOCA [16] employs a pre-trained automatic speech recognition (ASR) network to regress face vertex offsets from audio, achieving good lip-sync for multiple speaking styles. Many follow-up methods use a similar determinstic paradigm [3, 18, 24–26, 30, 33, 46, 57, 58, 60, 61, 75, 84, 93].

Stochastic neural methods. The first model to approach the problem stochastically was MeshTalk [69]. MeshTalk makes use of a pretrained discretized facial motion prior along with an explicit supervision mechanism that supports generation of motions that only have a weak correlation with the audio (eyebrow motion, etc.). MeshTalk employs autoregressive predicition and one can therefore sample the distribution of the next token to generate variety.

Thanks to the tremendous success of Diffusion Models in the image domain [36, 71] they are now widely used for 3D human body animation [9, 79], as well as speech-driven animation [73, 74, 83, 98]. The diffusion methods also employ transformer-based audio feature extractors [4, 38] to condition the denoiser. FaceDiffuser [73] employs a GRU-based denoiser, 3DiFACE [83] uses a TCN-based denoiser, and DiffPoseTalk [74], Media2Face [98] and FaceTalk [2] employ a transformer decoder architecture [86] passing the audio condition to the denoiser via cross-attention. Media2Face and DiffPoseTalk are the most similar to THUNDER— they use pseudo-GT on a large scale and utilize a transformer-decoder to denoise the animation in the 3DMM space.

Controlling the output animation. The task of controlling or editing the output animation is of great utility in production as manually editing animations is a laborious process. Many methods use a simple one-hot encoding of training subjects, which is mapped to a "style embedding" with a learnable layer. This embedding conditions the decoder to match the speaking style of the corresponding training subject [16, 26, 84, 91]. This paradigm has been extended to control the emotion and emotion intensity [18, 61]. DiffPoseTalk [74] uses contrastive learning to produce a style

vector from a reference animation, lifting the subject-ID conditioning limitation. Media2Face [98] employs CLIP features [66], which can be extracted from an image, or text prompts to condition the denoising process. 3DiFACE [83] demonstrates that the output of a diffusion model can be guided by a sparse set of keyframes. In this paper, instead of focusing on controlling the animation, we present a general method to improve lip-sync quality, which can be used in conjunction with the above control mechanisms.

Lip animation experts. A few methods attempt to design specific expert mechanisms to improve the lip-sync quality. EMOTE [18] uses an image-based lip-reading network on a differentiably-rendered output image in there, but the need for differentiable rendering of videos in-the-loop makes the approach slow and GPU-memory intensive. SelfTalk [60] proposes augmenting the FaceFormer architecture with a language decoder and language-based losses. A recent work by Chae et al. [8] employs masked auto-encoding and contrastive learning (akin to CLIP) to train a speech-mesh representation. The authors showed that the network can improve lip-animations in deterministic systems like FaceFormer or SelfTalk. To the best of our knowledge, no existing method attempts to synthesize audio for cycle-consistency, or employ lip animation experts in stochastic speech-driven animation.

Datasets. Most methods [16, 26, 45, 84, 91] have made use of high quality 3D scans synchronized with audio, such VOCASET [16] or Multiface [90] or Kinect-captured BIWI [27]. These datasets are expensive to acquire and hence are limited in number of subjects, and richness of facial expressions. As a consequence, even the stochastic methods that leverage these datasets [73, 91], suffer from lack of natural diversity of expressions. Thanks to the significant improvement of fast face reconstruction regressors [17, 29, 68, 97, 100], recent methods have turned to using pseudo-ground-truth (pGT) reconstructions from videos [18, 61, 74, 98]. While the pGT does not reach the quality of 3D scans, its quality is now sufficient. Video datasets are numerous, allowing more data to be acquired, which benefits data-hungry stochastic models.

2.2. Silent-video-to-speech (SVTS)

To the best of our knowledge, the prediction of voice from a sequence of 3D facial shapes has never been explored. However, recent years have seen tremendous progress on the task of speech audio prediction from silent videos. Early methods [15, 37, 47, 56, 59, 85, 88, 94] focus on small in-the-lab datasets with predefined scripts and limited number of speakers (GRID [14], TCD-TIMIT [34]) or script-unconstrained but single-speaker models [64].

SVTS [20] was the first method to produce intelligible audio on large-scale in-the-wild datasets such as LRS2 and LRS3 [1] by regressing spectrograms from mouth crops. The spectrograms are converted to the final waveform with a pretrained vocoder[12]. Follow-up methods improve the prediction quality [48] by adding a surrogate ASR loss or predicting features from an ASR model [12] The most recent architectures are based on diffusion, producing audio often indistinguishable from real speech [11, 96]. These, however, are not suitable for a self-supervised reconstruction loop, due to the iterative nature of diffusion models. Hence, thanks to its simple, yet efficient feed-forward design and a relatively straightforward architecture and training, we base our mesh-to-speech model on Choi et al.[12].

2.3. Audio cycle consistency

Cycle consistency losses on audio have been applied before, for tasks like voice conversion [42–44] and disentangled representation learning [10, 41]. Many different self-supervised audio learning methods exist (see survey [52]). However, to our knowledge, we are the first to propose cross-modal cycle consistency between audio and 3D facial motion.

3. Preliminaries

3.1. Face model

THUNDER uses the FLAME face model [51], which is a statistical 3D morphable model (3DMM). FLAME provides a compact representation of facial shapes and expressions and is defined as a function:

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}) \rightarrow (\mathbf{V}, \mathbf{F}), \tag{1}$$

where the inputs are shape coefficients $\boldsymbol{\beta} \in \mathbb{R}^{|\boldsymbol{\beta}|}$, expression coefficients $\boldsymbol{\psi} \in \mathbb{R}^{|\boldsymbol{\psi}|}$ and rotation vectors for $k = 4$ joints $\boldsymbol{\theta} \in \mathbb{R}^{3k+3}$. FLAME outputs a 3D mesh with vertices $\mathbf{V} \in \mathbb{R}^{n_v \times 3}$ and triangles $\mathbf{F} \in \mathbb{R}^{n_f \times 3}$. Since we do not focus on head movement, we only use the jaw joint rotation $\boldsymbol{\theta}_{jaw}$. For brevity, we refer to the expression coefficients and jaw rotation as expression parameters $\boldsymbol{x} = [\boldsymbol{\psi}|\boldsymbol{\theta}_{jaw}]$.

3.2. Speech feature extraction

Following previous work on speech-driven facial animation [18, 26, 84, 91, 98] we use Wav2Vec2.0 [4] as our audio feature extractor. It consists of a temporal convolution network (TCN) that extracts audio features at 50Hz and a transformer encoder that processes these features. Similarly to previous methods, we resample the TCN feature to 25Hz and feed it to the transformer to obtain the final speech feature. Formally, we define this as: $\mathcal{A}(\mathbf{w}) \rightarrow \mathbf{a}^{1:T}$, where $\mathcal{A}$ denotes the Wav2Vec 2.0 network, $\mathbf{w}$ is the input waveform, and $\mathbf{a}^{1:T} \in \mathbb{R}^{T \times d_s}$, with T denoting the number of frames at 25Hz and $d_s = 768$ is the feature dimension.

3.3. Speaker Embedding

A speaker embedding is a numerical representation of a speaker's unique vocal characteristics encoded into a vector. This vector captures features such as tone, pitch, speaking

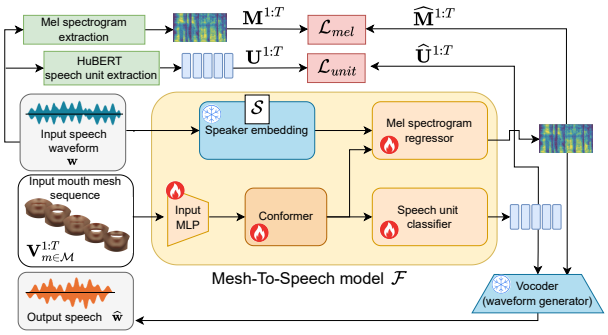

Figure 2. **Mesh-to-speech architecture.** It takes a sequence of mouth shapes as input, along with a speaker embedding feature, to produce the output speech units and spectrograms. These are used to compute a loss (top) and to produce the reconstructed audio (bottom) using a pretrained vocoder.

style, and other vocal attributes, enabling the differentiation and recognition of individual speakers. Among other applications, speaker embeddings are used in the latest SV2S models [12, 20, 48]. Following Choi et al. [12], we leverage an off-the-shelf speaker embedding extractor [39]. We define the extractor as: $\mathcal{S}(\mathbf{w}) \rightarrow \boldsymbol{s}$, with $\mathcal{S}$ and $\boldsymbol{s}$ denoting embedding extractor and the embedding vector, respectively.

3.4. Speech Units

The term speech units refers to discrete linguistic units identified through a self-supervised learning process. Like Choi et al. [12], we leverage speech units predicted by a pretrained HuBERT [38] model and its associated feature clustering model. First, features are extracted and subsequently clustered into $n_{\boldsymbol{U}}$ units. We define speech units as one-hot class vectors $\boldsymbol{U} \in [c_1, c_2, \ldots, c_{n_{\boldsymbol{U}}}]$ where only one of $c_i, i = 1 \ldots c_{n_{\boldsymbol{U}}}$ is 1 and the rest are 0. Speech units are predicted at 50Hz.

4. Method

THUNDER is trained in two stages. First, a mesh-to-speech (M2S) network is trained to regress audio from facial motion. Second, a diffusion model is trained to output 3D facial animation. The output animation is then fed to the frozen M2S model producing output audio representations. This allows us to design a loss between the input and output representations, which we utilize in training the talking head system.

4.1. Mesh-To-Speech

Regressing audio from 3D facial motion has, to our knowledge, never been done. To narrow the design space, we draw inspiration from the recent SV2S model by Choi et al. [12].

Architecture. The architecture consists of an input feature encoder, a conformer sequence encoder and two prediction heads, a speech unit classifier, and a mel spectrogram regressor. These two outputs contain enough information that a pretrained off-the-shelf vocoder [12] can turn them into output audio. Since the input is not a video, but a 3D animation, we replace the original lip video ResNet encoder that Choi et al. [12] use with an MLP that takes 3D lip vertex coordinates as the input. We keep the rest of the architecture the same. The M2S architecture is depicted in Fig. 2 and can be written

$$\mathcal{F}(\mathbf{V}_{m\in\mathcal{M}}^{1:T}, \boldsymbol{s}) \rightarrow (\widehat{\boldsymbol{M}}^{1:T}, \widehat{\boldsymbol{U}}^{1:T}), \tag{2}$$

where $\mathcal{F}(.)$ denotes the M2S network, $\mathcal{M}$ is a subset of mouth vertices. $\widehat{\boldsymbol{M}}^{1:T}, \widehat{\boldsymbol{U}}^{1:T}$ denote the output sequence of mel spectrograms and speech units. Finally, $\widehat{\boldsymbol{M}}^{1:T}$ and $\widehat{\boldsymbol{U}}^{1:T}$ can be passed to an off-the-shelf vocoder to generate the output waveform $\widehat{\mathbf{w}}$.

Supervision. We follow the same supervision scheme as Choi et al. [12]. The loss consists of two terms. The first term is an L1 loss between the input and output mel specrograms:

$$\mathcal{L}_{mel} = \|\boldsymbol{M}^{1:T} - \widehat{\boldsymbol{M}}^{1:T}\|_1. \tag{3}$$

The second term is cross entropy between the predicted speech unit classification vectors and the GT speech units:

$$\mathcal{L}_{unit} = \frac{1}{T}\sum_{t=1}^{T}\sum_{c=1}^{n_{\boldsymbol{U}}} \boldsymbol{U}_c^t \log p(\widehat{\boldsymbol{U}}_c^t), \tag{4}$$

with weights $w_{mel} = 10$ and $w_{unit} = 1$. The final loss is:

$$\mathcal{L}_{m2s} = w_{mel}\mathcal{L}_{mel} + w_{unit}\mathcal{L}_{unit}. \tag{5}$$

4.2. Speech-Driven Facial Animation Diffusion

Architecture. The core of THUNDER consists of a denoiser in the form of a transformer decoder [86]. Similar to other contemporary 3D animation methods (e.g. Zhao et al. [98] and Sun et al. [74]), THUNDER processes a noisy sequence of expression parameters $\tilde{\boldsymbol{x}}_{(d)}^{1:T}$, where d represents the diffusion timestep. These parameters serve as the *query* inputs to the transformer. The *keys* and *values* for the transformer are comprised of the input audio features $\mathbf{a}^{1:T}$ and a diffusion step embedding $\mathcal{T}(d)$, concatenated along the temporal dimension. The function $\mathcal{T}$ is defined as a sinusoidal timestep function similar to Ho et al. [36]. We do not incorporate additional animation-controlling conditions (such as style vectors [16, 26, 91], emotion vectors [18, 61] or CLIP embeddings [98]) as this is not the goal of this work. However, the design can easily be augmented with additional input conditions. All inputs are projected into the transformer's feature dimension, $f = 128$, using dedicated learnable affine layers. Positional encoding within the transformer's self-attention utilizes the ALiBi mechanism [65]. The cross-attention layers use a diagonal binary mask to facilitate the attention mechanism, and no additional positional encoding. The model predicts the reconstructed expression

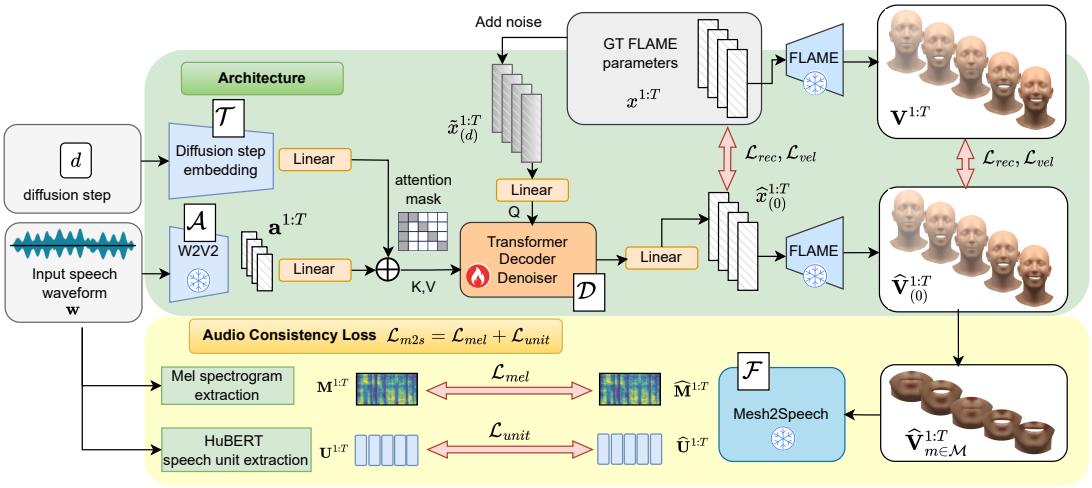

Figure 3. **THUNDER architecture.** The upper part of the figure (green) depicts the architecture of the diffusion model and the lower part (yellow) illustrates the application of the audio consistency loss. Gray boxes indicate the input to the system. Trainable components are highlighted in orange and frozen ones in blue.

parameters $\widehat{\boldsymbol{x}}_{(0)}^{1:T}$. The transformation can be formally described by the following equation:

$$\mathcal{D}(\tilde{\boldsymbol{x}}_{(d)}^{1:T}, d, \mathbf{a}^{1:T}) \rightarrow \widehat{\boldsymbol{x}}_{(0)}^{1:T}. \tag{6}$$

Here, $d \in \{1, \ldots, D\}$ indicates the diffusion timestep, and $\widehat{\boldsymbol{x}}_{(0)}^{1:T}$ is the denoised prediction. Unlike some methods that predict noise, this model directly predicts the fully denoised space, which is used to compute the 3D mesh sequence: $M(\boldsymbol{\beta} = \mathbf{0}, \widehat{\boldsymbol{x}}_{(0)}^{1:T}) \rightarrow \widehat{\mathbf{V}}_{(0)}^{1:T}$. This approach allows the model to define losses directly on the final 3D mesh space, crucial for leveraging M2S. See Fig. 3 for an overview.

Training. During training, a diffusion step d is sampled and $\tilde{\boldsymbol{x}}_{(d)}^{1:T}$ is computed from the clean expression parameters $\boldsymbol{x}^{1:T}$ using a noise schedule that blends the original data with Gaussian noise based on predefined α values:

$$\tilde{\boldsymbol{x}}_{(d)}^{1:T} = \sqrt{\alpha_d}\boldsymbol{x}^{1:T} + \sqrt{1-\alpha_d} \cdot \epsilon, \; \epsilon \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})), \tag{7}$$

where α_d represents the proportion of the original signal's variance preserved at diffusion timestep d, and ϵ is drawn from a Gaussian distribution with predefined $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. This approach incrementally corrupts the clean parameters, enabling the model to learn to denoise varying levels of noise. The model is supervised with several loss terms - reconstruction and velocity losses in expression space and 3D vertex space, and also the M2S loss (Eq. (5)). The reconstruction loss is given as:

$$\mathcal{L}_{rec} = \|\mathbf{V}^{1:T} - \widehat{\mathbf{V}}_{(0)}^{1:T}\|_2^2 + \|\boldsymbol{\psi}^{1:T} - \widehat{\boldsymbol{\psi}}_{(0)}^{1:T}\|_2^2 + \|\boldsymbol{\theta}_{jaw}^{1:T} - \widehat{\boldsymbol{\theta}}_{jaw_{(0)}}^{1:T}\|_2^2.$$

The velocity loss is also an MSE error but computed on the velocities of the individual components:

$$\mathcal{L}_{vel} = \|v(\mathbf{V}^{1:T}) - v(\widehat{\mathbf{V}}_{(0)}^{1:T})\|_2^2 + \|v(\boldsymbol{\psi}^{1:T}) - v(\widehat{\boldsymbol{\psi}}_{(0)}^{1:T})\|_2^2 + \|v(\boldsymbol{\theta}_{jaw}^{1:T}) - v(\widehat{\boldsymbol{\theta}}_{jaw_{(0)}}^{1:T})\|_2^2,$$

where v computes the velocity of any input sequence $\mathbf{u}^{1:T}$ $v(\mathbf{u}^{1:T}) = \mathbf{u}^{2:T} - \mathbf{u}^{1:T-1}$. The final loss is given as:

$$\mathcal{L}_{total} = w_{m2s}\mathcal{L}_{m2s} + \mathcal{L}_{rec} + \mathcal{L}_{vel}, \tag{8}$$

with $w_{m2s} = 1$. The model is trained using classifier-free guidance [35]. The audio condition is randomly dropped and replaced with a learnable vector with 20% probability.

Inference. At inference time, we initialize the first step by sampling the Gaussian noise: $\widehat{\boldsymbol{x}}_{(D)} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We then proceed to denoise for D diffusion steps. In each step, we first employ the denoiser $\mathcal{D}(\tilde{\boldsymbol{x}}_{(d)}^{1:T}, d, \mathbf{a}^{1:T})$ to obtain $\widehat{\boldsymbol{x}}_{(0)}^{1:T}$. Then, we compute $\tilde{\boldsymbol{x}}_{(d-1)}$ by adding the expected amount of noise back to obtain $\tilde{\boldsymbol{x}}_{(d-1)}$. Since THUNDER is trained with classifier-free guidance, it is possible to trade-off fidelity of the lip-sync for increased diversity by combining $s_a\mathcal{D}(\tilde{\boldsymbol{x}}_{(d)}^{1:T}, d, \mathbf{a}^{1:T}) + (1 - s_a)\mathcal{D}(\tilde{\boldsymbol{x}}_{(d)}^{1:T}, d, \emptyset)$ during the denoising process. However, we find our results are diverse enough without this and so we effectively set $s_a = 1$.

5. Experiments

5.1. Datasets

The selection of datasets for THUNDER is challenging. 4D scan datasets such as VOCASET [16] provide good GT but are too small to train generative models. In order to have a high-enough quality geometry from which speech can be predicted, we use a collection of public in-the-lab captured video datasets commonly used for lip-reading, namely TCD-TIMIT [34], RAVDESS [53] and GRID [14], totaling 117 subjects, uttering 42140 short sentences, totaling 3535816 frames at 25FPS (over 39 hours). These datasets provide clear imagery without in-the-wild complexity and can be reconstructed by off-the-shelf methods to a satisfactory degree.

Following EMOTE [18], we employ the publicly available INFERNO framework [19], which contains a SOTA face reconstruction system [17, 29, 100]. We include an expanded discussion on face tracker selection in Sec. 9 in Sup. Mat. We split each of the datasets by subjects, 70% for training and 15% each for validation and testing. We refer to this data as THUNDERSET. Additionally, we also train and test M2S and THUNDER on a more challenging TFHP [74], which contains sequences in-the-wild and head pose.

5.2. Mesh-To-Speech

Implementation details. We initialize the conformer and prediction heads with pretrained weights from Choi et al. [12], while the input MLP is trained from scratch. We train the model with the Adam optimizer [49] for 20 epochs and select the checkpoint with the lowest validation loss (usually reached in 5 epochs). The training batch size is 16 and the input sequence is trimmed to 125 frames at 25FPS.

Quantitative comparison. Since the M2S model is the first of its kind, we compare M2S to a recent SV2S model [12] which we finetune on the videos from our training dataset for fair comparison. Note that FLAME does *not* model the teeth or tongue, which makes the M2S task arguably harder than SV2S. Teeth and tongue, which may be visible on video, are critical for production of certain phonemes, such as labiodentals (/f/, /v/), or alveolar phonemes (/s/, /z/, /t/, /d/, /l/) [50]. Consequently, we do not expect M2S to match the performance of SV2S methods.

Table 1 reports the standard video-to-speech metrics, namely Short-Time Objective Intelligibility [76], Extended STOI [40] to measure the intelligibility of the generated samples, Perceptual Evaluation of Speech Quality (PESQ) [70] (narrow and wide band) to measure the perceptual quality, and Word Error Rate (WER), which we measure with the SOTA ASR network Whisper[67]. As expected, M2S is not as accurate as video-to-speech, but the performance is comparable, suggesting that there is sufficient audio signal in the 3D to be useful for our downstream task. Note that M2S is, however, more accurate than Choi et al.'s original (but not the fine-tuned) model. We experiment with the following input spaces for M2S: selection of mouth vertices $\mathbf{V}_{m\in\mathcal{M}}$ (*mouth2s*), all vertices $\mathbf{V}$ (*face2s*), and FLAME expression parameters $\boldsymbol{x}$ (*exp2s*) and report the results in Tab. 1. Remarkably, the model which takes global expression vectors, performs slightly better than the selection of mouth vertices $\mathbf{V}_{m\in\mathcal{M}}$. Despite that, we choose *mouth2s* to be our final model, as it results in better lip-sync supervision (discussed in Sec. 5.3 and Tab. 2). Please refer to the Sup. Video for audible comparison and Sec. 7.1 in Sup. Mat. for further discussion on performance.

Modality	Model:	STOI ↑	ESTOI ↑	PESQ-WB ↑	PESQ-NB ↑	WER ↓
Mesh-to-speech (M2S)	exp2speech $\boldsymbol{x}$	0.502	*0.273*	*1.257*	*1.468*	*0.648*
	face2speech $\mathbf{V}$	0.458	0.203	1.225	1.437	0.953
	mouth2speech $\mathbf{V}_{m\in\mathcal{M}}$	*0.506*	0.272	1.246	1.457	0.678
Video-to-speech (SV2S)	Choi et al. (finetuned)	**0.555**	**0.348**	**1.281**	**1.511**	**0.437**
	Choi et al. (orig)	0.376	0.141	1.126	1.313	1.011

Table 1. **Quantitative comparison** of video-to-speech and forms of M2S (*mouth2speech*, *face2speech* and *exp2speech*).

5.3. Speech-Driven Facial Animation

Implementation details. THUNDER is trained using the Adam optimizer [49] ($lr = 1e-4$) for 260 epochs with batch size 48, sequence length of 70 frames, and $D = 1000$ diffusion steps with a linear noise schedule. The transformer has 8 layers with 4 heads and feature dim $f = 128$.

Baselines. We compare THUNDER to other speech-driven avatar methods that are also trained on pGT and which output 3DMM parameters (as opposed to vertices). We reimplemented and trained two SOTA methods, namely FlameFormer* and FlameSelfTalk* (adaptations of FaceFormer [26] and SelfTalk [60]), Media2Face* [98] and DiffPoseTalk* [74] (see Sup. Mat. for details). Unless stated otherwise, we keep the pretrained Wav2Vec2 weights frozen. Models that finetune it are denoted with suffix -T (T for trainable). Reimplemented methods are marked with asterisks*.

Evaluation protocol. For each test audio sequence, we generate S=32 outputs. Each of these is initialized with a different random noise in the case of diffusion models, or randomly sampled subject-id condition in case of FlameFormer. Then we compute the following evaluation metrics, averaged out across the S outputs (unless stated otherwise).

Lip Vertex Error. Following [69, 74, 91] we report LVE, which calculates the maximum L2 error of all lip vertices for each frame and then computes the average over all frames.

Dynamic Time Warping. We also report the DTW proposed by Thambiraja et al. [84]. First, distances between the midpoints of lower and upper lips are calculated for both the predictions and GT. The two resulting time series are then used to compute the DTW distance.

Lip Correlation Coefficients. LVE is limited to sparse vertices and ignores temporal dynamics, while DTW reduces lip motion to a single distance. To complement these, we treat each mouth-region vertex coordinate as a time series and compute Pearson (synchrony) and Concordance (synchrony and bias/scale agreement) correlation coefficients between pGT and predictions.

Face dynamic deviation. Following CodeTalker [91], we report Face Dynamics Deviation (FDD), which measures the difference between the temporal std. deviation of the pGT and the predicted vertices, averaged over a set of vertices. We report upper-FDD (FDD-U) and lip-FDD (FDD-L).

Sample diversity. One of the most important aspects of stochastic models is that they generate diverse animations for

Experiment	Input	Model	Lip-Sync				Upper-face Diversity [cm]		Lip Diversity [cm]		Face Dynamic Dev. [cm]	
			LVE [cm] ↓	L-CCC ↑	L-PCC ↑	DTW [cm] ↓	S-DIV-U ↑	T-DIV-U ↑	S-DIV-L ↓	T-DIV-L ↓	T-FDD-U ↓	T -FDD-L ↓
(1) THUNDER and mesh-to-speech input space	audio only	THUNDER w/o m2s	0.879	0.359	0.568	0.329	**0.0419**	**0.044**	0.21	0.254	*0.0118*	0.0932
		THUNDER w/ face2s	*0.804*	*0.411*	*0.633*	**0.285**	0.0297	*0.0409*	**0.128**	*0.237*	**0.0117**	*0.0827*
		THUNDER w/ exp2s	0.83	0.362	0.63	0.296	*0.0404*	0.0398	0.176	**0.228**	0.0125	0.0939
		THUNDER w/ mouth2s	**0.802**	**0.426**	**0.639**	*0.29*	0.0322	0.04	*0.134*	0.241	0.0122	**0.0806**
(2) THUNDER and trainable audio enc.	audio only	THUNDER-T w/o m2s	0.723	0.428	0.623	0.266	0.020	**0.0411**	**0.0656**	0.216	0.0118	0.0811
		THUNDER-T w/ mouth2s	**0.709**	**0.445**	**0.66**	**0.256**	**0.021**	0.039	0.0669	**0.202**	0.0118	**0.0788**
(3) Mesh-to-speech with other diffusion-based methods which use additional input conditions	audio, image feature	Media2Face*	*0.96*	*0.308*	*0.531*	*0.363*	**0.0163**	*0.0459*	*0.108*	*0.256*	**0.0105**	*0.0899*
		Media2Face* w/ m2s	**0.815**	**0.428**	**0.609**	**0.282**	*0.0128*	**0.0586**	**0.0371**	**0.233**	*0.0154*	**0.083**
	audio, speaker style feat.	DiffPoseTalk*	*0.68*	*0.464*	*0.612*	**0.267**	**0.025**	**0.0379**	*0.117*	*0.223*	**0.0101**	**0.0776**
		DiffPoseTalk* w/ m2s	**0.651**	**0.469**	**0.653**	*0.268*	*0.0172*	*0.0377*	**0.0604**	**0.204**	*0.0117*	*0.0813*
(4) Mesh-to-speech with deterministic methods and discrete input conditions	audio, one-hot speaker ID	FlameFormer*	*0.809*	*0.368*	*0.57*	*0.291*	**0.0271**	*0.0372*	*0.132*	*0.239*	*0.0117*	*0.0909*
		FlameFormer* w/ m2s	**0.794**	**0.411**	**0.614**	**0.265**	*0.0263*	**0.0393**	**0.111**	**0.219**	**0.0107**	**0.0715**
	audio, one-hot speaker ID	FlameSelfTalk*	*0.695*	*0.504*	*0.698*	**0.243**	*0.0275*	*0.0372*	*0.108*	*0.201*	*0.00983*	*0.0692*
		FlameSelfTalk* w/ m2s	**0.674**	**0.518**	**0.705**	*0.243*	**0.029**	**0.039**	**0.108**	**0.193**	**0.00965**	**0.0674**

Table 2. **Quantitative evaluation** on THUNDERSET **(1) Mesh-to-speech input space.** All of the M2S variants are applicable. THUNDER with *mouth2s* results in the best lip-sync performance (LVE, L-CCC, L-PCC), which is why we select it as the final THUNDER model. *exp2s* is applicable, too as it scores best on DTW and has good diversity (S-DIV-U), while *face2s* scores the worst on diversity (S-DIV-U, T-DIV-U). **(2) Trainable audio encoder.** THUNDER-T results in improved lip-sync metrics (LVE, L-CCC, L-PCC and DTW), but at slight expense of diversity of outputs (mainly upper face diversity S-DIV-U and T-DIV-U) compared to THUNDER which does not train the audio encoder. **(3) Mesh-To-Speech with diffusion-based methods.** Training Media2Face* (with image conditioning) and DiffPoseTalk* (style conditioning) follows the same trend - increase in lip-sync quality and reduction in upper-face diversity. **(4) Mesh-To-Speech with deterministic methods.** We train FlameFormer* and SelfTalk* with M2S and find that it improves the lip-sync related metrics.

the same input. Remarkably, previous work has not analyzed this aspect [73, 74, 98]. We compute standard deviations of vertex distances between the pGT and the predictions for the same input at frame t. Formally, given a tensor of per-vertex L2 distances $\mathbf{D} \in \mathbb{R}^{N \times S \times T \times n_v}$, with N being the size of the test dataset and $S = 32$ is the number of sampled outputs per input. We compute the standard deviation along the sample dimension S. Other dimensions are subsequently averaged. We measure sample diversity for the upper face region (S-Div-U) and the lip region (S-Div-L).

Temporal diversity. We also report the temporal diversity, where the standard deviation is computed over the temporal dimension of $\mathbf{D} \in \mathbb{R}^{N \times S \times T \times n_v}$ (and averaged across the rest). Again, we calculate the diversity for both lip (T-Div-L) and upper face regions (T-Div-U).

5.3.1 Quantitative evaluation

Tab. 2 reports the test-set metrics computed for models trained on THUNDERSET and Tab. 3 does so for models trained on TFHP. We analyze the following:

Does M2S improve THUNDER? Yes, models with M2S improve all lip-sync metrics at slight expense of diversity (see Tab. 2, experiments 1 and 2).

Does the input space for M2S matter? While all models improve lip-sync metrics, *mouth2speech* has the strongest effect on lip-sync metrics, likely thanks to its localized effect on the mouth region (Tab. 2, experiment 1).

Should the audio encoder be finetuned? Trainable encoder (what most methods use) results in much lower diversity, presumably due to overfitting to the training voices. Frozen encoder results in inferior lip-sync but higher diversity. M2S improves lip-sync in both cases. M2S applied with a frozen encoder leads to a dramatically improved lip-sync while retaining some of the upper-face diversity (Tab. 2, exp. 1,2).

Does M2S improve other methods? Yes, we apply M2S to train the deterministic FlameFormer, SelfTalk and stochastic diffusion-based Media2Face and DiffPoseTalk and show improvement on the lip-sync metrics (Tab. 2, exp. 3,4) .

Does THUNDER generalize to other datasets? Yes. We train and test M2S and THUNDER on the challenging TFHP (the original DiffPoseTalk dataset). THUNDER outperforms original DiffPoseTalk (see Tab. 3 and Sec. 7.2.1 in Sup. Mat.)

Can THUNDER generate head pose motion? Yes. We train a THUNDER model with head pose on TFHP and show improvement over DiffPoseTalk (see Tab. 3 and Sec. 7.2.1).

Does M2S help with editing methods? Yes. We feed images of different emotions to our M2F. M2F w/ M2S improves lip-sync (see Sec. 7.2.2 in Sup. Mat.)

Perceptual study. We conduct a perceptual study in which we evaluate THUNDER and the effect of mesh-to-speech. Specifically, we compare to the deterministic FlameFormer, THUNDER with and without M2S, and p-GT. The participants are shown two videos side-by-side and asked to rate three aspects of the (lip-sync, dynamism, and realism) ani-

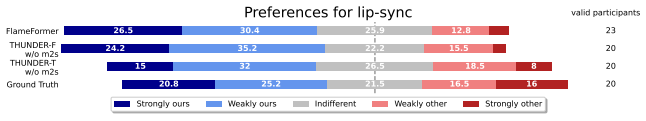

Figure 4. **Perceptual study of THUNDER.** We compare THUNDER-F (-F for frozen backbone) with methods having both a trainable audio encoder (THUNDER-T w/o m2s) and frozen encoders (FlameFormer-F, THUNDER-F w/o m2s), and GT. The participants prefer THUNDER-F's lip-sync over that of the other models. Remarkably, the participants also have slight preference for THUNDER-F over GT, which suggests that the application of M2S helps THUNDER saturate the quality of GT.

		Lip-Sync				Upper-face diversity [mm]		Lip diversity [mm]		Face Dynamic Dev. [mm]		Head Pose	
Name	Backbone	LVE [cm] ↓	L-CCC ↑	L-PCC ↑	DTW [cm] ↓	S-DIV-U ↑	T-DIV-U ↑	S-DIV-L ↓	T-DIV-L ↓	FDD-U ↓	FDD-L ↓	BA $\times 10^{-3}$ ↑	DIV ↑
THUNDER-F w/o m2s	Wav2Vec2 frozen	*1.2*	*0.285*	*0.388*	*2.26*	**0.258**	**0.349**	*2.6*	*3.58*	**0.0649**	**0.76**	*6.02*	*3.07*
THUNDER-F	Wav2Vec2 frozen	**1.07**	**0.298**	**0.435**	**1.95**	*0.203*	*0.321*	**1.5**	**2.98**	*0.0707*	*0.95*	**6.13**	**3.78**
DiffPoseTalk	HuBERT trainable	*1.01*	*0.414*	*0.541*	*1.85*	**0.247**	**0.352**	*1.58*	*3.02*	0.0641	*0.618*	5.54	2.16
THUNDER-T w/o m2s	Wav2Vec2 trainable	1.12	0.358	0.488	*1.86*	0.241	*0.351*	1.87	3.21	*0.0597*	0.661	**6.25**	*3.49*
THUNDER-T	Wav2Vec2 trainable	**0.987**	**0.422**	**0.55**	**1.81**	*0.244*	0.341	**1.35**	**2.89**	**0.0578**	**0.55**	*5.77*	**4.83**

Table 3. **Quantitative evaluation on TFHP.** We compare THUNDER, and THUNDER w/o M2S and the original DiffPoseTalk release (all trained on TFHP). Consistently to our other experiments, THUNDER results in superior lip-sync performance over THUNDER w/o M2S and DiffPoseTalk, at a slight trade-off of diversity metrics. Notably, when trained with head pose, all THUNDER variants outperforms DiffPoseTalk on head pose metrics - beat alignment (BA) and head pose diversity (DIV). Refer to Sup. Video and PDF for qualitative results.

mations on a five-point Likert scale (strong/weak preference for A or B or indifferent). We find that participants prefer THUNDER over other methods. Fig. 4 reports the results for THUNDER for lip-sync. For the remaining results of the study and further discussion, see Sec. 7.2.4 in Sup. Mat.

5.3.2 Qualitative evaluation

We refer the reader to the supplemental video for a detail comparisons of the animations, which demonstrates the superiority of our method qualitatively. Fig. 5 shows the comparison of our method with the baselines for selected utterances. Additional qualitative evaluation can be found in Sup. Mat..

5.4. Limitations

THUNDER has the following limitations. The absence of teeth and tongue inherently creates ambiguity, making it more difficult to produce high quality audio compared to silent video-to-speech. Further, pseudo-GT is not a perfect reconstruction but has occasional artifacts or inaccuracies. Also, our current M2S architecture is fully deterministic and hence is not completely capable of capturing the prediction ambiguities. All of these may result in suboptimal effects of the audio cycle-consistency loss. Regardless of these shortcomings, our experiments demonstrate the benefits of our method. Employing a large-scale dataset of high quality 3D scans, a face model that models teeth and tongue, and upgrade M2S to a stochastic model are likely to boost the benefits of the M2S loss. Finally, the inference process of the diffusion model is currently computationally intensive. Future work should address this by employing more efficient solvers such as DPM++ [54].

6. Conclusion

In this paper, we introduced THUNDER, a 3D speech-driven avatar generation system with a novel paradigm of supervision via analysis-by-audio-synthesis. THUNDER has two key components. First, drawing inspiration from silent video-to-speech methods, we define a new mesh-to-speech task and train a model that regresses speech from facial animation. Second, we incorporated mesh-to-speech into a diffusion-based 3D speech-driven avatar system, creating the first-of-its-kind audio-based self-supervision loop that

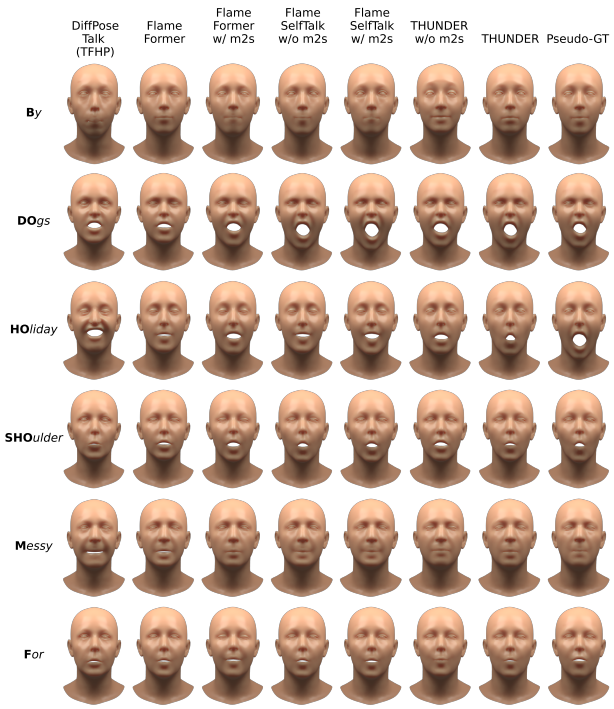

Figure 5. **Qualitative comparison on THUNDERSET.** This figure shows the comparison between baselines, our model and GT for selected utterances. Note that DiffPoseTalk was trained on TFHP. Supplemental PDF and video contain more qualitative comparisons.

helps ensure that the produced lip animation is plausible for the spoken audio. Our extensive quantitative, perceptual, and qualitative experiments demonstrate that THUNDER achieves a significant improvement in lip animation quality for diffusion-based 3D speech-driven avatars while still producing a diverse set of facial animations. Furthermore, we have demonstrated that the application of our mesh-to-speech loss improves the lip-sync quality of other talking head avatar systems. Finally, we believe that our novel audio-based self-supervision paradigm can impact other applications of 3D head avatars, such as video-based reconstruction, automatic postprocessing of 3D facial animation, or as a quality measure of talking head avatars.

Acknowledgments. We thank Raja Bala, Hiro Takeda, Brandon Smith, Ganesh Iyer, Chinghang Chen, Alex Vorobiov and Betty Mohler for helpful discussions. This project was supported by an Amazon Research Award.

Disclosure: While MJB is a co-founder and Chief Scientist at Meshcapade, his contribution was performed at, and funded by, the MPG.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition. *CoRR*, abs/1809.00496, 2018. 3, 13

[2] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. Facetalk: Audio-driven motion diffusion for neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21263–21273, 2024. 2

[3] Shivangi Aneja, Artem Sevastopolsky, Tobias Kirschstein, Justus Thies, Angela Dai, and Matthias Nießner. Gaussianspeech: Audio-driven personalized 3d gaussian avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13065–13075, 2025. 2

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020. 2, 3

[5] Volker Blanz, Sami Romdhani, and Thomas Vetter. Face identification across different poses and illuminations with a 3D morphable model. pages 202–207, 2002. 12

[6] Volker Blanz, Curzio Basso, Tomaso A. Poggio, and Thomas Vetter. Reanimating faces in images and video. *Computer Graphics Forum (Proc. EUROGRAPHICS)*, 22(3):641–650, 2003. 12

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

[8] Lee Chae-Yeon, Oh Hyun-Bin, Han EunGi, Kim Sung-Bin, Suekyeong Nam, and Tae-Hyun Oh. Perceptually accurate 3d talking head generation: New definitions, speech-mesh representation, and evaluation metrics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21065–21074, 2025. 3, 2

[9] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 2

[10] Kiran Chhatre, Radek Daněček, Nikos Athanasiou, Giorgio Becherini, Christopher Peters, Michael J. Black, and Timo Bolkart. AMUSE: Emotional speech-driven 3D body animation via disentangled latent diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1942–1953, 2024. 3

[11] Jeongsoo Choi, Joanna Hong, and Yong Man Ro. Diffv2s: Diffusion-based video-to-speech synthesis with vision-guided speaker embedding. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 7778–7787. IEEE, 2023. 3

[12] Jeongsoo Choi, Minsu Kim, and Yong Man Ro. Intelligible lip-to-speech synthesis with speech units. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 4349–4353. ISCA, 2023. 2, 3, 4, 6

[13] Michael M. Cohen, Rashid Clark, and Dominic W. Massaro. Animated speech: research progress and applications. In *Auditory-Visual Speech Processing, AVSP 2001, Aalborg, Denmark, September 7-9, 2001*, page 200. ISCA, 2001. 2

[14] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. The grid audio-visual speech corpus, 2006. 3, 5

[15] Thomas Le Cornu and Ben Milner. Reconstructing intelligible audio speech from visual speech features. In *16th Annual Conference of the International Speech Communication Association, INTERSPEECH 2015, Dresden, Germany, September 6-10, 2015*, pages 3355–3359. ISCA, 2015. 3

[16] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. Capture, learning, and synthesis of 3d speaking styles. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10101–10111. Computer Vision Foundation / IEEE, 2019. 1, 2, 3, 4, 5

[17] Radek Danecek, Michael J. Black, and Timo Bolkart. EMOCA: emotion driven monocular face capture and animation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 20279–20290. IEEE, 2022. 3, 6, 12

[18] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*, New York, NY, USA, 2023. Association for Computing Machinery. 1, 2, 3, 4, 6, 12, 13

[19] Radek Daněček. INFERNO: Set the world on fire with FLAME. https://github.com/radekd91/inferno, 2023. 6, 13

[20] Rodrigo Schoburg Carrillo de Mira, Alexandros Haliassos, Stavros Petridis, Björn W. Schuller, and Maja Pantic. SVTS: scalable video-to-speech synthesis. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 1836–1840. ISCA, 2022. 3, 4

[21] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 285–295. Computer Vision Foundation / IEEE, 2019. 12

[22] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: An animator-centric viseme model for expressive lip synchronization. *ACM Trans. Graph.*, 35(4), 2016. 2

[23] Pif Edwards, Chris Landreth, Mateusz Popławski, Robert Malinowski, Sarah Watling, Eugene Fiume, and Karan Singh. Jali-driven expressive facial animation and multilingual speech in cyberpunk 2077. In *ACM SIGGRAPH 2020 Talks*, New York, NY, USA, 2020. Association for Computing Machinery. 2

[24] Han EunGi, Oh Hyun-Bin, Kim Sung-Bin, Corentin Nivelet Etcheberry, Suekyeong Nam, Janghoon Ju, and Tae-Hyun

Oh. Enhancing speech-driven 3d facial animation with audio-visual guidance from lip reading expert. 2024. 2

[25] Xiangyu Fan, Jiaqi Li, Zhiqian Lin, Weiye Xiao, and Lei Yang. Unitalker: Scaling up audio-driven 3d facial animation through A unified model. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLI*, pages 204–221. Springer, 2024.

[26] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18749–18758. IEEE, 2022. 1, 2, 3, 4, 6, 12

[27] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6):591 – 598, 2010. 3

[28] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. 40(4):88:1–88:13, 2021. 12

[29] Panagiotis P. Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Visual speech-aware perceptual 3d facial expression reconstruction from videos, 2022. 3, 6, 1, 4, 12, 13

[30] Hui Fu, Zeqing Wang, Ke Gong, Keze Wang, Tianshui Chen, Haojie Li, Haifeng Zeng, and Wenxiong Kang. Mimic: Speaking style disentanglement for speech-driven 3d facial animation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 1770–1777. AAAI Press, 2024. 2

[31] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Pixel3dmm: Versatile screen-space priors for single-image 3d face reconstruction, 2025. 12

[32] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 15180–15190. IEEE, 2023. 7

[33] Kazi Injamamul Haque and Zerrin Yumak. Facexhubert: Text-less speech-driven e(x)pressive 3d facial animation synthesis using self-supervised speech representation learning. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23)*, New York, NY, USA, 2023. ACM. 2

[34] Naomi Harte and Eoin Gillen. TCD-TIMIT: an audio-visual corpus of continuous speech. *IEEE Trans. Multim.*, 17(5): 603–615, 2015. 3, 5

[35] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. 5

[36] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2, 4

[37] Joanna Hong, Minsu Kim, Se Jin Park, and Yong Man Ro. Speech reconstruction with reminiscent sound via visual voice memory. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3654–3667, 2021. 3

[38] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. 2, 4

[39] Corentin Jemine. Real-time voice cloning. https://github.com/CorentinJ/Real-Time-Voice-Cloning, 2023. 4

[40] Jesper Jensen and Cees H. Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE ACM Trans. Audio Speech Lang. Process.*, 24(11):2009–2022, 2016. 6

[41] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14080–14089. Computer Vision Foundation / IEEE, 2021. 3

[42] Takuhiro Kaneko and Hirokazu Kameoka. Parallel-data-free voice conversion using cycle-consistent adversarial networks. *CoRR*, abs/1711.11293, 2017. 3

[43] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 6820–6824. IEEE, 2019.

[44] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. Cyclegan-vc3: Examining and improving cyclegan-vcs for mel-spectrogram conversion. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, pages 2017–2021. ISCA, 2020. 3

[45] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 1, 2, 3

[46] Hyung Kyu Kim, Sangmin Lee, and Hak Gu Kim. Memorytalker: Personalized speech-driven 3d facial animation via audio-guided stylization. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11241–11251, 2025. 2

[47] Minsu Kim, Joanna Hong, and Yong Man Ro. Lip to speech synthesis with visual context attentional GAN. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 2758–2770, 2021. 2, 3

[48] Minsu Kim, Joanna Hong, and Yong Man Ro. Lip-to-speech synthesis in the wild with multi-task learning. In *ICASSP*

2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE, 2023. 2, 3, 4

[49] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[50] Peter Ladefoged and Keith Johnson. A course in phonetics (sixth edition). 2011. 6

[51] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. 36(6):194:1–194:17, 2017. 3, 12

[52] Shuo Liu, Adria Mallol-Ragolta, Emilia Parada-Cabaleiro, Kun Qian, Xin Jing, Alexander Kathan, Bin Hu, and Björn W. Schuller. Audio self-supervised learning: A survey. *Patterns*, 3(12):100616, 2022. 3

[53] Steven R. Livingstone and Frank A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), 2018. 5

[54] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Mach. Intell. Res.*, 22(4):730–751, 2025. 8

[55] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172, 2019. 12

[56] Daniel Michelsanti, Olga Slizovskaia, Gloria Haro, Emilia Gómez, Zheng-Hua Tan, and Jesper Jensen. Vocoder-based speech synthesis from silent videos. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, pages 3530–3534. ISCA, 2020. 3

[57] Federico Nocentini, Thomas Besnier, Claudio Ferrari, Sylvain Arguillère, Stefano Berretti, and Mohamed Daoudi. Scantalk: 3d talking heads from unregistered scans. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXIX*, pages 19–36. Springer, 2024. 2

[58] Federico Nocentini, Claudio Ferrari, and Stefano Berretti. Emovoca: Speech-driven emotional 3d talking heads. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2859–2868. IEEE, 2025. 2

[59] Dan Oneata, Adriana Stan, and Horia Cucu. Speaker disentanglement in video-to-speech conversion. In *29th European Signal Processing Conference, EUSIPCO 2021, Dublin, Ireland, August 23-27, 2021*, pages 46–50. IEEE, 2021. 3

[60] Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 5292–5301. ACM, 2023. 2, 3, 6

[61] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. *CoRR*, abs/2303.11089, 2023. 1, 2, 3, 4

[62] Hai Xuan Pham, Samuel Cheung, and Vladimir Pavlovic. Speech-driven 3d facial animation with implicit emotional awareness: A deep learning approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2328–2336. IEEE Computer Society, 2017.

[63] Hai Xuan Pham, Yuting Wang, and Vladimir Pavlovic. End-to-end learning for 3d facial animation from raw waveforms of speech. *CoRR*, abs/1710.00920, 2017. 2

[64] K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 13793–13802. Computer Vision Foundation / IEEE, 2020. 3

[65] Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 4

[66] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 3, 2

[67] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 28492–28518. PMLR, 2023. 6

[68] George Retsinas, Panagiotis P. Filntisis, Radek Danecek, Victoria F. Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2490–2501, 2024. 3, 12, 13

[69] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1153–1162. IEEE, 2021. 2, 6

[70] Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2001, 7-11 May, 2001, Salt Palace Convention Center, Salt Lake City, Utah, USA, Proceedings*, pages 749–752. IEEE, 2001. 6

[71] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. 2

[72] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J. Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7763–7772. Computer Vision Foundation / IEEE, 2019. 12

[73] 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. 2, 3, 7

[74] 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. *ACM Trans. Graph.*, 43(4):46:1–46:9, 2024. 2, 3, 4, 6, 7, 1, 12, 13

[75] Kim Sung-Bin, Lee Chae-Yeon, Gihun Son, Oh Hyun-Bin, Janghoon Ju, Suekyeong Nam, and Tae-Hyun Oh. Multitalk: Enhancing 3d talking head generation across languages with multilingual video dataset. In *25th Annual Conference of the International Speech Communication Association, Interspeech 2024, Kos, Greece, September 1-5, 2024*. ISCA, 2024. 2

[76] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. Speech Audio Process.*, 19(7):2125–2136, 2011. 6

[77] Sarah L. Taylor, Moshe Mahler, Barry-John Theobald, and Iain A. Matthews. Dynamic units of visual speech. In *Proceedings of the 2012 Eurographics/ACM SIGGRAPH Symposium on Computer Animation, SCA 2012, Lausanne, Switzerland, 2012*, pages 275–284. Eurographics Association, 2012. 2

[78] Sarah L. Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica K. Hodgins, and Iain A. Matthews. A deep learning approach for generalized speech animation. *ACM Trans. Graph.*, 36(4): 93:1–93:11, 2017. 2

[79] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[80] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Pérez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3735–3744. IEEE Computer Society, 2017. 12

[81] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2549–2559. Computer Vision Foundation / IEEE Computer Society, 2018.

[82] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. FML: face model learning from videos. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10812–10822. Computer Vision Foundation / IEEE, 2019. 12

[83] Balamurugan Thambiraja, Sadegh Aliakbarian, Darren Cosker, and Justus Thies. 3diface: Diffusion-based speech-driven 3d facial animation and editing, 2023. 2, 3

[84] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation, 2023. 2, 3, 6

[85] Seyun Um, Jihyun Kim, Jihyun Lee, and Hong-Goo Kang. Facetron: A multi-speaker face-to-speech model based on cross-modal latent representations. In *31st European Signal Processing Conference, EUSIPCO 2023, Helsinki, Finland, September 4-8, 2023*, pages 281–285. IEEE, 2023. 3

[86] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4

[87] Thomas Vetter and Volker Blanz. Estimating coloured 3D face models from single images: An example based approach. In *ECCV*, pages 499–513, 1998. 12

[88] Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Video-driven speech reconstruction using generative adversarial networks. In *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, pages 4125–4129. ISCA, 2019. 3

[89] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. MEAD: A large-scale audio-visual dataset for emotional talking-face generation. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXI*, pages 700–717. Springer, 2020. 13

[90] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason M. Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. Multiface: A dataset for neural face rendering. 2022. 3

[91] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven

3d facial animation with discrete motion prior. pages 12780–12790, 2023. 2, 3, 4, 6

[92] Yuyu Xu, Andrew W. Feng, Stacy Marsella, and Ari Shapiro. A practical and configurable lip sync method for games. In *Motion in Games, MIG '13, Dublin, Ireland, November 6-8, 2013*, pages 131–140. ACM, 2013. 2

[93] Zhihao Xu, Shengjie Gong, Jiapeng Tang, Lingyu Liang, Yining Huang, Haojie Li, and Shuangping Huang. Kmtalk: Speech-driven 3d facial animation with key motion embedding. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LVI*, pages 236–253. Springer, 2024. 2

[94] Ravindra Yadav, Ashish Sardana, Vinay P. Namboodiri, and Rajesh M. Hegde. Speech prediction in silent videos using variational autoencoders. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 7048–7052. IEEE, 2021. 3

[95] Karren D. Yang, Anurag Ranjan, Jen-Hao Rick Chang, Raviteja Vemulapalli, and Oncel Tuzel. Probabilistic speech-driven 3d facial motion synthesis: New benchmarks, methods, and applications. pages 27284–27293, 2024. 2, 12, 13

[96] Yochai Yemini, Aviv Shamsian, Lior Bracha, Sharon Gannot, and Ethan Fetaya. Lipvoicer: Generating speech from silent videos guided by lip reading. 2024. 3

[97] Tianke Zhang, Xuangeng Chu, Yunfei Liu, Lijian Lin, Zhendong Yang, Zhengzhuo Xu, Chengkun Cao, Fei Yu, Changyin Zhou, Chun Yuan, and Yu Li. Accurate 3d face reconstruction with facial component tokens. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 8999–9008. IEEE, 2023. 3, 12

[98] Qingcheng Zhao, Pengyu Long, Qixuan Zhang, Dafei Qin, Han Liang, Longwen Zhang, Yingliang Zhang, Jingyi Yu, and Lan Xu. Media2face: Co-speech facial animation generation with multi-modality guidance, 2024. 2, 3, 4, 6, 7, 12

[99] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audio-driven animator-centric speech animation. *ACM Trans. Graph.*, 37(4), 2018. 2

[100] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIII*, pages 250–269. Springer, 2022. 3, 6, 1, 4, 12, 13