

## EmoVOCA: Speech-Driven Emotional 3D Talking Heads

Federico Nocentini

University of Florence, Italy

federico.nocentini@unifi.it

Claudio Ferrari

University of Parma, Italy

claudio.ferrari@unipr.it

Stefano Berretti

University of Florence, Italy

stefano.berretti@unifi.it

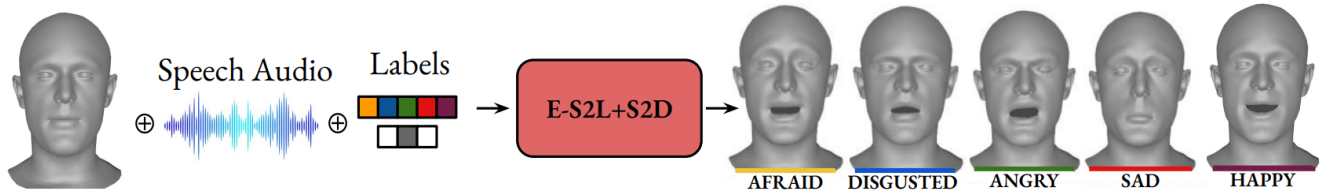


Figure 1. We introduce **EmoVOCA**, a novel approach for generating a synthetic 3D Emotional Talking Heads dataset which leverages speech tracks, intensity labels, emotion labels, and actor specifications. The proposed dataset can be used to surpass the lack of 3D datasets of expressive speech, and train more accurate emotional 3D talking head generators as compared to methods relying on 2D data as proxy.

### Abstract

A notable challenge in 3D talking head generation consists in blending speech-related motions with expression dynamics. This is primarily caused by the lack of comprehensive 3D datasets that combine diversity in spoken sentences with a variety of facial expressions. Some literature works attempted to overcome such lack of data by fitting parametric 3D models (3DMMs) to 2D videos, and using the reconstructed 3D faces as replacement. However, their underlying parametric space limits the precision required to accurately reproduce convincing lip motions and syncing, which is crucial for the application at hand. In this work, we look at the problem from a different perspective, and developed a data-driven technique to combine inexpressive 3D talking heads with a set of 3D expressive sequences, which we used for creating a synthetic dataset, called **EmoVOCA**. We then designed and trained an emotional 3D talking head generator that accepts a 3D face, an audio file, an emotion label, and an intensity value as inputs, and learns to animate the audio-synchronized lip movements with expressive traits of the face. Comprehensive experiments, both quantitative and qualitative, using our data and generator evidence superior ability in synthesizing convincing animations, when compared with the best performing methods in the literature. Our code and pre-trained models are available at <https://github.com/miccunifi/EmoVOCA>.

### 1. Introduction

Generating 3D talking heads from speech aims at animating a 3D facial model with dynamic lip movements that correspond to a spoken sentence. While current state-of-

the-art methodologies such as [6, 9, 15–17, 19–21, 25, 28] effectively replicate lip-syncing and facial deformations, achieving natural emotional expression in animated faces lacks full conviction. The primary constraint inhibiting research progress in emotional 3D talking heads resides in the absence of suitable 3D datasets that allow learning the complex interaction between geometrical mouth deformations induced by speech and expressions. In fact, directly collecting such data is a rather unpractical solution due to costly devices and a time-consuming capturing process. A workaround that was explored recently consists in leveraging 2D video datasets as a proxy. For example, Peng *et al.* [21] and Daněček *et al.* [6] extracted per-frame 3D faces from video datasets of emotional speech thanks to parametric 3D models (3DMMs). In doing so, they gathered sequences of plausible 3D faces paired with the speech extracted from the videos. A shortcoming of these approaches, though, is that the precision and variety of the 3D estimated facial deformations are bounded by the underlying parametric model, with a consequent loss of subtle lip movements. In fact, while 3DMMs are good for modeling facial shapes and expressions, they are not sufficiently expressive for reproducing speech-related lip motions.

In the attempt of solving the above, we explore a different approach and propose to combine an existing 3D talking head dataset, where captured subjects only show a neutral emotional state (VOCAsset [4]), with 3D emotional faces from a dataset of 3D expressive sequences (Florence4D [22]). We chose these specific datasets as the meshes they include share the same topology. The major challenge here consists in realistically combining speech-related mouth motions with expression-induced facial deformations. To simplify the problem, some methods [23]

consider an additional constraint, and assume that facial deformations resulting from expressions are isolated in the upper half of the face, forcing a spatial disentanglement of such motions. While this solution can work for natural facial movements, the above assumption only partially holds true when emotions are involved. Suffice it to say, anger or happiness induce changes in the mouth shape, which in turn influence its motion during speech. We here tackle the challenge of combining speech and expression induced facial deformations from 3D face datasets and surpass the need for parametric 3D models, in order to build training data for emotion-conditioned 3D talking head generators. To this aim, we propose a framework composed of two encoders and a single, shared decoder. The idea is that each encoder learns speech or expression specific features, while the decoder learns both. Once trained, we can combine speech and expression related motions by simple feature combination, and let the decoder generate the entangled deformations. To showcase the applicability of our synthesized data, we used them to train a network that generates emotional 3D talking heads from a speech track, an intensity label, an emotion label, and an actor to animate, as illustrated in Fig. 1. To this end, we developed on two state-of-the-art solutions [9, 17] that generate unexpressive talking heads based on the sole audio features, and adjusted them to be conditioned with an emotion and an intensity label. In summary the main contributions of our work are:

- We introduce a new data-driven approach to realistically *entangle* speech and expression specific facial deformations, providing more accurate training data for building expressive talking head generators;
- With the above, we synthesize EmoVOCA, a new 3D dataset combining speech with emotional nuances;
- On such data, drawing from state-of-the-art techniques, we designed and trained two deep architectures that generate accurate expressive 3D talking heads using an audio, an emotion, and intensity labels;
- In an extensive set of experiments, we unveil limitations of previous methods, and the advantages of our scheme with respect to prior works.

## 2. Related Work

First efforts primarily concentrated on the synthesis of facial animation through the manipulation of pre-defined facial rigs using procedural mechanisms. Subsequent to this, facial animations were constructed through viseme-dependent co-articulation models or via the blending of facial templates [7, 8, 10]. These methods showcased a synthesis framework that combines language and facial expressions at a fundamental level. Karras *et al.* [11] harnessed

a dataset-driven approach, learning a 3D facial animation model from a limited but high-quality 3D dataset of a specific actor. This strategy stands in contrast to VOCA [4], which leveraged a more expansive dataset featuring diverse subjects, capable of animating an array of corresponding identities from audio cues. *MeshTalk* [23] proposed a complementary solution to those discussed above by learning a categorical representation for facial expressions. Sampling in an auto-regressive way from this categorical space, this approach can animate a given 3D facial template mesh of a subject from audio inputs. *FaceFormer* [9] proposed a Transformer-based autoregressive model, which encodes the long-term audio context and autoregressively predicts a sequence of animated 3D face meshes. Nocentini *et al.* [17], aimed to address the challenges encountered by FaceFormer and VOCA, like the long training time, and increased the lip-movements accuracy, introducing S2L+S2D. This framework enhances the lip-sync abilities of talking heads by guiding them through a landmark-based motion paradigm. In the most recent effort to develop speech-driven 3D talking heads, Stan *et al.* proposed *FaceDiffuser* [25], an approach centered around a Gated Recurrent Unit model. This model was trained to operate like a diffusion model, predicting the 3D face from a set of Gaussian noise inputs. We also mention the work by Thambiraja *et al.* [27], who proposed *Imitator* to add personalized traits to the talking head.

All the aforementioned methods were trained on publicly available datasets. However, these datasets pose a limitation as the recorded faces lack expressions, resulting in a deficiency of emotional content in the generated faces. To address the lack of specific data, researchers have tested various techniques. Lu *et al.* [14] proposed to use EMOCAv2 [5] to reconstruct 3D talking heads from in-the-wild videos. Chang *et al.* [3] proposed a fusion of 2D and 3D datasets to inject emotional information into the generated faces. Conversely, Peng *et al.* [21] devised a methodology to reconstruct 3D emotional talking head datasets from 2D sources based on a facial blendshape capturing method. Upon constructing the dataset, they developed a Transformer-based model, called *Emotalk*, for generating 3D emotional talking heads. Differently from others, they did not condition the generation with emotion labels, but directly extracted emotional context from the audio. More recently, Sung-Bin *et al.* [26] proposed LaughTalk, a framework proficient in crafting speech-driven 3D laughing talking heads. Additionally, Daněček *et al.* [6] presented EMOTE, a framework capable of generating emotional 3D talking heads predicated on audio input and emotion labels.

Again, the very core of these approaches is based upon the efficacy of reconstructing a 3D face from 2D representations. We will show that, this strategy results in decreased lip-sync accuracy.

### 3. Proposed Approach

The lack of publicly available emotional 3D talking head datasets motivated our efforts. We propose here a solution to address most of the limitations faced by previous works that attempted to exploit 2D video data as an intermediate means to gather 3D expressive talking heads. To this aim, we leverage two publicly available 3D face datasets: one includes 3D sequences of talking faces in *neutral expression*; the other, instead, comprises sequences of 3D faces portraying *emotion related facial expressions*, yet without speech-related lip movements. No additional assumptions are imposed on the data; for instance, sequences are not required to have uniform lengths, nor is synchronization necessary between instances in the two datasets.

Our idea is to generate emotional 3D talking heads by learning to explicitly *entangle* speech and emotion face deformations by mixing latent codes that are separately learned from the two datasets. The framework, illustrated in Fig. 2, is composed of two encoders, which separately learn to embed speech or emotion related facial deformations, and a shared decoder. The model is trained to reconstruct the input deformations. Being the decoder shared across the two datasets, at inference time the embeddings can be combined so to generate the entangled deformations. Ultimately, this allowed us to create a new dataset, called EmoVOCA, that can be employed for training emotional 3D talking heads generation approaches.

#### 3.1. Data Preparation

We utilized two 3D face datasets, referred to as  $D_T$  and  $D_E$ :  $D_T$  contains *inexpressive* 3D talking heads, while  $D_E$  consists of *emotional* 3D sequences without speech-driven lip activation. Both  $D_T$  and  $D_E$  are in FLAME [12] topology. This choice simplified the framework design by allowing us to use the same encoder architecture, but does not represent a strict constraint. The two encoders do not share weights and are trained separately (Sec. 3.2), so they can be adapted to process 3D faces of different topologies. In order to learn speaker-independent speech or expression deformations, we first pre-process the data to remove the identity component from each face. This is obtained by subtracting, for each subject, the face in neutral expression, *i.e.*, with no expression or lip movement, from each instance of the corresponding subject in both datasets. Based on this, the datasets  $D_T$  and  $D_E$  are represented as:

$$D_T = \{(M_i^t, M_i^n)\}_{i=0}^P, \quad D_E = \{(M_j^e, M_j^n)\}_{j=0}^H, \quad (1)$$

where,  $M_i^t$  and  $M_j^e$  represent, respectively, the  $i$ -th talking head and the  $j$ -th emotional head, while  $M_i^n$ ,  $M_j^n$  denote the corresponding neutral configurations.  $P$  and  $H$  are the number of samples in the  $D_T$  and  $D_E$  datasets, respectively.

For each face  $M_i^t \in D_T$  and  $M_j^e \in D_E$ , the *dense deformation offset* is computed as the difference between the animated face and its neutral configuration:

$$S_i^t = M_i^t - M_i^n, \quad S_j^e = M_j^e - M_j^n. \quad (2)$$

This processing step generates a displacement-based representation encoding speech and expression motions from the two source datasets  $S_T = \{S_i^t\}_{i=0}^P$ , and  $S_E = \{S_j^e\}_{j=0}^H$ . Though both datasets contain sequences, the frames are treated individually. Thus, there is no need for the length of the sequences to match.

#### 3.2. Double Encoder/Shared Decoder Architecture

Our proposed architecture is made up of a Double-Encoder and a Shared-Decoder (DE-SD), and is summarized in Fig. 2. The building block for both encoders and decoder is the *SpiralNet* proposed by Bouritsas *et al.* [2]. The two distinct encoders independently process speech or expression related data:  $E_T$  uses samples in  $S_T$  to learn capturing speech-related movements, while  $E_E$  processes samples in  $S_E$  to learn deformations due to expressions. Both encoders embed the samples into separate latent vectors:  $f_i^t = E_T(S_i^t)$  and  $f_j^e = E_E(S_j^e)$ . However, these feature vectors eventually need to be combined so that the decoder can learn both motions and generate mixed deformations. Given that no ground-truth exists for the combined motions, how the features are concatenated depends on whether we are training or testing the model.

##### 3.2.1 Training Phase

Ground-truth samples are available only for each dataset separately, thus each encoder is trained alternatively. However, we need to maintain a consistent embedding size at both training and inference. Hence, the embedded feature vectors are duplicated and concatenated during training. These are then fed to the decoder to reconstruct the input:  $\hat{S}_i^t = D(f_i^t \oplus f_i^t)$  for  $E_T$ , and  $\hat{S}_j^e = D(f_j^e \oplus f_j^e)$  for  $E_E$ . This strategy allows us to later combine features from both encoders. The training objective is to reconstruct the input displacements. We used a weighted  $L_2$  loss:

$$L = \frac{1}{N} \sum_{i=1}^N w_i \left\| S_i - \hat{S}_i \right\|_2. \quad (3)$$

where  $N$  represents the number of vertices in the meshes, while  $S_i$  and  $\hat{S}_i$  are respectively, the ground-truth and predicted displacements at vertex  $i$ . Here,  $w_i$  represents the per-vertex weight as introduced by Otterdout *et al.* [18] that measures the contribution of each vertex in the mesh according to the inverse of its distance from the closest landmark. This factor penalizes points far from movable face areas, *i.e.*, mouth, eyes, as they remain mostly stationary and so do not contribute to modeling facial deformations.

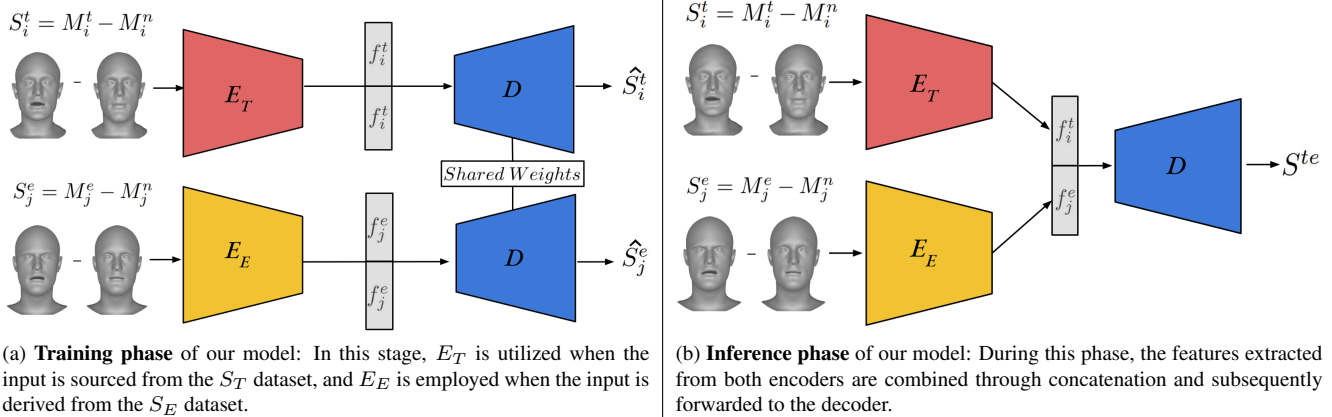


Figure 2. Overview of our framework. (a) Two distinct encoders, namely  $E_T$  and  $E_E$  process the talking and expressive 3D head displacements, separately, while a common decoder  $D$  is trained to reconstruct them. (b) At inference, talking and emotional heads are combined by concatenating the encoded latent vectors, and the decoder outputs a combination of their displacements.

### 3.2.2 Inference Phase

Given talking displacements  $S_i^t$ , and emotional displacements  $S_j^e$ , we use the encoders to embed them, deriving  $f_i^t = E_T(S_i^t)$  and  $f_j^e = E_E(S_j^e)$ . These feature vectors are then concatenated and fed to the decoder to generate the mixed motion:  $S^{te} = D(\mu_t f_i^t \oplus \mu_e f_j^e)$ . Here,  $\mu_t$  and  $\mu_e$  are coefficients that allow us to adjust the contribution of the two sets of features. By modifying these coefficients, we can control the interplay between the talking and emotional displacements, significantly augmenting the span of the generation variety. Once we obtain  $S^{te}$ , we can add these displacements to a neutral face and obtain the 3D emotional talking head as  $M_i^{te} = M_i^n + S_i^{te}$ .

### 3.2.3 On the Architectural Design

Features in the two datasets are disentangled as speech and expression deformations involve *different face regions*, so their spatial distribution overlaps only to some extent. Thus, to correctly reconstruct the displacements (either  $S^t$  or  $S^e$ ) during training, the decoder needs to output spatially different, almost complementary, displacement fields. Our intuition to duplicate the features is that the decoder is made up of spiral convolution layers, which operate locally, and are based on the knowledge of the mesh graph [2]. Thus each “neuron” in each layer corresponds to a vertex in the (sub-sampled) mesh. Since the two deformations are spatially disentangled, speech embeddings  $f^t$  activate different decoder features with respect to emotion embeddings  $f^e$  so that there is no need to explicitly factorize the latent space. Even if features are duplicated during training, the prior spatial disentanglement forces the decoder to activate the same set of neurons. When features are combined, it instead induces the activation of all neurons, leading to realistic entanglement of the deformation (see Fig. 3c).

## 4. Experimental Results

In the following, we first introduce the datasets and the metrics (Sec. 4.1), then we summarize the methods for emotional talking heads generation that we trained on the data synthesized with our DE-SD architecture (Sec. 4.2). We report quantitative and qualitative results plus a test involving the proposed methods in Sec. 4.3, and Sec. 4.4, respectively. Finally, in Sec. 4.6, we conducted two user studies to compare our approach with current SOTA methods for emotional talking heads, *i.e.*, Emotalk [21] and EMOTE [6].

### 4.1. Datasets and Metrics

We utilized two datasets, VOCaset and Florence 4D, to train our DE-SD network, which was subsequently used to generate the EmoVOCA dataset.

**VOCaset** [4] contains 3D talking head sequences from 12 actors, evenly split between 6 males and 6 females. Each actor recorded 40 sentences, with durations ranging from 3 to 5 seconds. The dataset provides per-frame 3D facial reconstructions captured at 60 frames per second (fps) along with corresponding audio recordings. However, VOCaset lacks head movements and upper facial expressions.

**Florence 4D** [22] offers 3D dynamic sequences representing 70 different emotions and facial expressions. Each sequence consists of 60 meshes, with the peak facial expression typically occurring between frames 25 and 35.

**EmoVOCAv1** is a collection of expressive 3D talking head sequences generated using our DE-SD architecture. It combines features from VOCaset and Florence 4D to generate sequences for five emotions: *afraid*, *angry*, *disgust*, *happy*, and *sad*, each with three intensity levels by varying the  $\mu_e$  factor. This produced a total of 7,200 3D sequences.

**EmoVOCAv2** expands upon EmoVOCAv1, adding *moody*, *drunk*, *ill*, *suspicious*, *pleased*, and *upset*, for a total of 11 emotions, with three intensity levels, resulting in 15,840 se-

quences. All datasets were divided into training (8 actors), validation, and test sets (2 actors each).

Following previous works [9, 21, 25], we used the following metrics, measured in millimeters:

- **MVE** (Max Vertex Error): The maximum  $L_2$  error between the predicted and ground truth mesh vertices, providing a global measure of reconstruction accuracy.
- **UVE** (Upper Vertex Error): The maximum  $L_2$  error computed on the vertices in the eyes and forehead regions, focusing on emotion-related movements.
- **LVE** (Lip Vertex Error): The maximum  $L_2$  error computed on the vertices around the mouth region, capturing both speech-/emotion-related movements.

MVE was used as a global metric to evaluate the overall reconstruction quality of DE-SD, as shown in Tab. 1. For talking head evaluation, UVE and LVE were chosen as local descriptors to assess the quality of expression-related and speech-related movements, respectively, as shown in Tab. 3.

## 4.2. Emotional 3D Talking Head Generators

We used both EmoVOCAv1 and EmoVOCAv2 to train two properly customized state-of-the-art solutions for 3D talking heads generation, namely Faceformer [9], and S2L+S2D [17]. These methods are constrained to generate neutral talking heads, so we adapted them for training and testing on our data. The primary goal is to show that by training on EmoVOCAv1/v2 we can both (i) surpass the accuracy of SOTA methods trained on the original VOCAs, and (ii) perform better than state-of-the-art emotional talking head generators that exploit 2D video data and parametric head models. Both the models were conditioned using audio features extracted from the wav2vec 2.0 [1] encoder, which was shown to be highly effective [6, 9, 17, 21, 27]. To add the emotion information and condition the generation, we represented each emotion label  $L$ , and intensity label  $I$ , as a one-hot vector, and embedded them as 64-dimensional vectors using a linear layer.

**Emo(E)-Faceformer** is inspired by Faceformer [9]. The architecture proposed by Fan *et al.* involves a transformer decoder, trained autoregressively, conditioned with audio features. We modified the original architecture by concatenating the two embeddings  $L$ ,  $I$  resulting from the emotion and intensity labels. Similar to Faceformer, we trained the model autoregressively with the same masking and the same periodic positional encoding. This method learns to animate a 3D face given an audio input, an emotion label and an intensity label. The  $L_2$  loss between the prediction and the ground truth was used for training.

**Emo(E)-S2L+S2D** develops upon the S2L+S2D method in [17]. This method works by decoupling the generation

in two steps: first, a bi-directional LSTM model (S2L) generates the motion of 68 landmarks of the face from the audio signal. Then, a mesh decoder (S2D) expands the landmarks motion to all the vertices of the face mesh. We trained the E-S2L with the landmarks extracted from the meshes, then we separately trained the S2D decoder to densify the movement of landmarks into the face mesh vertices. Similar to E-FaceFormer, we concatenated the two embeddings  $L$ ,  $I$  to the audio features. The E-S2L learns to predict landmarks movement based on an audio input, an emotion label, and an intensity label. To train the E-S2L model, we used the average  $L_2$  loss between the prediction and the ground truth. To train the S2D, we used the weighted  $L_2$  defined in (3).

## 4.3. EmoVOCA Evaluation

**Quantitative results.** We first assess the accuracy of our DE-SD model to reconstruct the test sets of VOCAs and Florence 4D. In Tab. 1, we report the results of an ablation study aimed at evaluating the effect of the weighted  $L_2$  loss applied during training and a different design for DE-SD trained with a single, shared encoder.

Table 1. Ablation study. Contributions of: weighted loss in [18] (left); double or single encoder (right). Errors are in mm.

	VOCAs		Florence 4D	
<b>Loss</b>	<b>MVE ↓</b>	<b>MVE ↓</b>	<b>MVE ↓</b>	<b>MVE ↓</b>
Standard $L_2$	0.931	0.844	0.745	0.676
<b>Weighted <math>L_2</math></b>	<b>0.722</b>	<b>0.657</b>	<b>0.722</b>	<b>0.657</b>

	VOCAs		Florence 4D	
<b>Encoder</b>	<b>MVE ↓</b>	<b>MVE ↓</b>	<b>MVE ↓</b>	<b>MVE ↓</b>
Single	0.745	0.676	0.722	0.657
<b>Double</b>	<b>0.722</b>	<b>0.657</b>	<b>0.722</b>	<b>0.657</b>

We first investigate the effect of the weighted loss [18]. Predictably, the latter leads to significantly improved outcomes as the model better focuses on movable facial areas. Another significant result is the improvement achieved by employing a double-encoder. This outcome aligns with our expectations, considering that each encoder specializes in extracting facial features that are either speech or emotion-related, further pushing a disentangled feature representation. This specialization arises from the datasets difference: in VOCAs, the upper face remains stable, whereas in Florence 4D it plays a more pivotal role (Sect. 3.2.3).

To further assess the feature concatenation strategy of DE-SD, we trained a simple expression classifier (Spiral Convolution Encoder with an MLP Head) using the Florence4D mesh sequences (expression-only). We then tested it on five different versions of our EmoVOCA dataset (expression+speech) generated using five different feature combination strategies in DE-SD. This was done to evaluate how well expressions are preserved when combined with speech. We trained DE-SD using different strategies: (i) summing (Sum) or multiplying (Mult) the embeddings  $f^t$  and  $f^e$ ; (ii) concatenating neutral (zero) vectors instead of duplicating the embeddings; and (iii) testing the original DE-SD by reversing the positions of  $f^t$  and  $f^e$  to verify the intuitions discussed in Sec. 3.2.3. The results for each

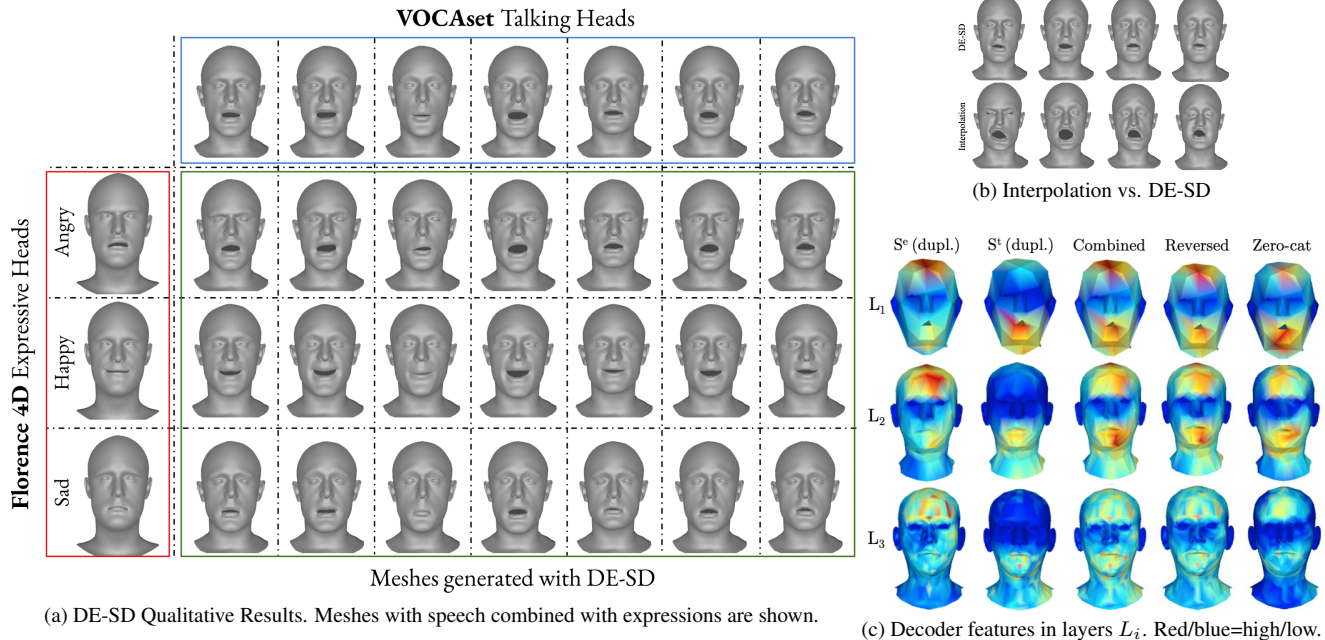


Figure 3. **(a)** The blue box (top row) shows some meshes from VOCaset, while the red box (left column) showcases various expressive heads from Florence 4D. In the green box, the face at row  $i$  and column  $j$  is the result of the combination of the  $i$ -th expressive and  $j$ -th talking heads, *i.e.*,  $S^{te} = D(E_T(S_j^t) \oplus E_E(S_i^e))$ . So, each column includes the talking head on the top conditioned with different expressions. **(b)** Comparison of DE-SD (top row) with displacement interpolation (bottom row). Simple interpolation in 3D space leads to unrealistic deformations and loss of speech accuracy. **(c)** Comparison of several strategies of feature combination. The decoder features are shown for each layer  $L_i$ . Even if features are duplicated during training, the prior spatial disentanglement forces the decoder to activate different neurons. When combined, the features induce a combined action of all neurons, leading to realistic entanglement.

intensity level are shown in Tab. 2. We also compared the geometric similarity of the generated meshes against those in the original datasets to determine which method best preserves facial features. Specifically, we utilized the LVE for VOCaset and the UVE for Florence 4D. Both complementary measures prove that duplicating features during DE-SD training leads to more recognizable expressions, even when the embeddings were reversed at inference, outperforming the other strategies.

Table 2. Classification accuracy for different expr. intensities. Error on both training datasets.

Model	$I_1$ Acc	$I_2$ Acc	$I_3$ Acc	LVE VOCaset ↓	UVE Florence 4D ↓
DE-SD Sum	0.40	0.64	0.72	6.181	1.073
DE-SD Mult	0.33	0.55	0.66	6.897	1.248
DE-SD Zero-cat	0.37	0.58	0.69	6.324	1.137
DE-SD Reversed	0.45	0.68	0.76	5.872	<b>0.940</b>
DE-SD	<b>0.46</b>	<b>0.71</b>	<b>0.82</b>	<b>5.861</b>	0.981

**Qualitative Results.** Fig. 3a shows various expressive meshes combined with talking heads with DE-SD. The deformation of the upper face is particularly influenced by the expressive mesh, although the mouth shape is influenced by the expression as well, even if to a lesser extent. This unbalanced influence is valuable as it infuses expressiveness to the talking heads, at the same time preserving the lip movements that correspond to the spoken sentence. Using two

separate encoders led to an effective disentangled learning of the motion priors, leading to a strong compositional ability [24]. This is instead compromised in previous methods such as EmoTalk or EMOTE (see Sec. 4.6).

It could be argued that the same goal could be achieved by simply linearly combining the speech and expression offsets ( $S_i^t$  and  $S_i^e$ ) or by morphing the faces using some Morphable Model parameters. As depicted in Fig. 3b, this is not the case. Displacement interpolation introduces severe artifacts, and leads to the loss of speech-related movements, resulting in lower realism in the generated faces. By directly summing the expressive offsets  $S_i^e$  to inexpressive talking heads generated by FaceFormer or S2L-S2D, we also quantitatively show that this is not a viable solution either (see results in Tab. 3a). Finally, examples in Fig. 3c qualitatively support the technical choice expounded in Sect. 3.2.3 and the results in Tab. 2. The average activation of decoder features are shown for each layer  $L_i$ . Duplicating either  $f^t$  or  $f^e$  activate distinct decoder features and, when combined, all features activate, contributing to a realistic entanglement. Reversing their arrangement does not have a significant influence, meaning that the two encoders effectively captured the disentangled deformations. Forcing an explicit latent arrangement by concatenating a zero vector would lose this ability, also resulting in samples of poorer quality.

#### 4.4. Generating Emotional 3D Talking Heads

Both E-Faceformer and E-S2L-S2D models were trained using separately both EmoVOCAv1 and EmoVOCAv2. We report average generation results across the test set of both versions in Tab. 3a. Based on the results presented in Tab. 3, it is evident that training a model using EmoVOCAv2 yields improved performance compared to training with EmoVOCAv1. This enhancement is a result of the increased number of samples in EmoVOCAv2. The E-S2L-S2D method clearly performs favorably with respect to E-FaceFormer (consistently with the result reported in [17]).

To conduct a more comprehensive evaluation of our generators E-Faceformer and E-S2L+S2D, we performed additional tests using the (inexpressive) VOCA-Test as suggested by [21]. We set the one-hot encoding of both emotion and intensity to zero, for both E-Faceformer and E-S2L+S2D. Results in Tab. 3b show that our emotional talking head generators exhibit good generalization ability to speech-related movements, surpassing the performance of the same architectures trained on the original VOCAset. Moreover, E-S2L+S2D yields superior performance compared to both EmoTalk and EMOTE. These outcomes further highlight the good quality of the samples synthesized with our DE-SD, suggesting our solution as a viable way to create complex yet effective 3D datasets.

Table 3. Quantitative comparison with previous works: (a) test on EmoVOCAv1 and EmoVOCAv2; (b) tests on VOCA-Test. (v1) and (v2) indicate training on EmoVOCAv1 and EmoVOCAv2, respectively; (Voca) indicates training on VOCAset; (2D) indicates training on 3D datasets reconstructed from videos.

Method	LVE (mm) ↓	UVE (mm) ↓
VOCA(Voca)	6.993	1.201
Faceformer(Voca)	6.123	1.117
SelfTalk(Voca)	5.618	0.989
S2L+S2D(Voca)	4.789	0.934
FaceDiffuser(Voca)	4.350	0.925
CodeTalker(Voca)	3.651	0.912
EmoTalk(2D)	4.134*	-
EMOTE(2D)	4.561	0.897
E-Faceformer(v1)	4.371	0.964
E-S2L+S2D(v1)	3.697	0.867
E-Faceformer(v2)	3.789	0.912
E-S2L+S2D(v2)	<b>3.181</b>	<b>0.791</b>

(a) Results from the baselines. The first two rows are tested on EmoVOCAv1.

(b) Results on VOCA-Test dataset. For the EmoTalk (\*) model, we report the results as presented in the original paper.

For a more detailed analysis, in Fig. 4 we report results for all the emotions and intensity levels separately. We observe a rather stable accuracy across the different emotions and intensities, although the UVE measure grows for those involving stronger movements of the upper face, such as *angry* or intensity *high*. In Fig. 5, we show qualitative examples also for EmoTalk and EMOTE (quantitative comparison was not possible, see Sec. 4.6). Note that our samples and the EMOTE ones are generated performing a zero-shot test using an audio from RAVDESS [13]. This was needed because EmoTalk only infers the emotion from the audio, and the VOCAset includes only neutral speech. Qualitatively, the results for E-S2L-S2D are the most realistic both

in terms of speech mimicking and perceived emotion (see also Tab. 4 and Tab. 5). Notably, EmoTalk can convey some emotional content, yet only by slightly moving the upper face area, whereas the mouth region remains rather static. This results both in smaller emotional effects, and reduced speech realism. EMOTE, instead, can generate good and recognizable emotional content, yet it is less accurate in terms of lip synchronization (see Tab. 5).

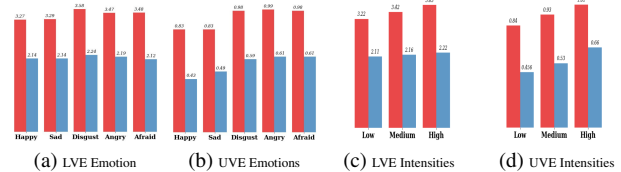


Figure 4. Generation results comparison for each emotion and each intensity level. The blue bars refer to E-S2L-S2D, while red bars to E-FaceFormer. Both baselines are trained and tested on EmoVOCAv1.

#### 4.5. E-S2L+S2D Additional Features

The E-S2L+S2D model, trained with EmoVOCA, efficiently generates emotional talking heads based on one-hot encoding of emotion and intensity labels. While training activates only one emotion and one intensity level at a time, the model’s high generalization enables diverse capabilities during inference. It can produce sequences with temporal emotion or intensity transitions, varying emotion and intensity labels during generation. In Fig. 6, we can see several examples of emotion and intensity transition by varying the one-hot encodings in a chosen moment in time.

These capabilities emerge from both the model’s design and the EmoVOCA dataset used for training. Consequently, a model trained on EmoVOCA can generate sequence of faces with emotion and intensity variation despite being trained on separate tasks and not encountering such scenarios during training. We highlight this a unique feature that is not shown by any other method or dataset. Qualitative examples can be found in the supplementary material.

#### 4.6. User based Study

To our knowledge, EmoTalk [21] and EMOTE [6] are the only approaches in recent literature for generating emotional 3D talking heads. Quantitative comparison is not possible due to discrepancies in mesh topology and dataset utilization across EmoTalk and EMOTE. EmoTalk’s mesh topology diverges from FLAME, while EMOTE’s training dataset differs from ours, impeding direct evaluation. Our decision to not re-implementing EmoTalk is based on its similarity to FaceFormer, employing a transformer decoder, with emotion inferred from audio. Similarly, re-training EMOTE with our dataset was not feasible due to its reliance on 2D video-based loss computation, which we

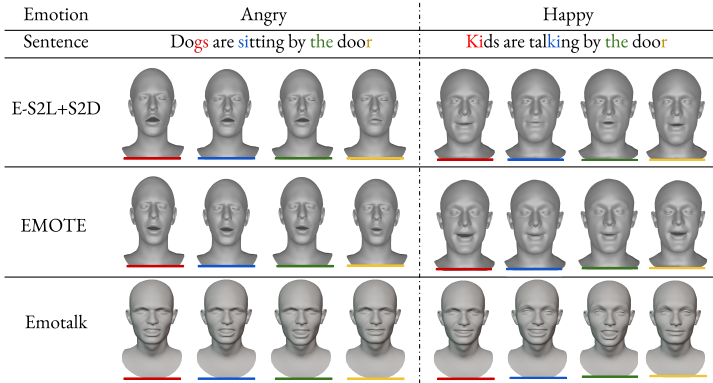


Figure 5. Qualitative examples generated by E-S2L+S2D trained on EmoVOCAv1 in comparison with EmoTalk and EMOTE. Colors represent specific moments in time.

lack. Consequently, we conducted two user-based studies involving 22 and 20 participants, spanning expert and non-expert, to assess our E-S2L+S2D trained on EmoVOCAv1 in comparison with EmoTalk and EMOTE. Following methodologies in prior works [9, 21, 27], we designed A/B studies for user evaluation. Thirty videos, generated using E-S2L+S2D trained on EmoVOCAv1, EmoTalk and EMOTE, were divided as follows: (i) Ten videos generated from RAVDESS [13] test set audios, asking users to assess lip synchronization fidelity and emotion portrayal (Test 1); (ii) Ten videos generated from *in-the-wild* YouTube audios, asking evaluations on lip synchronization (Test 2); (iii) Ten videos from RAVDESS audios were presented to users *without audio*, asking them to classify the emotion the generated videos seem to portray, so to verify if users can infer the emotion looking at the animation (Test 3).

Results from both user studies, summarized in Tab. 4 and Tab. 5, clearly favors E-S2L+S2D trained on EmoVOCAv1 for its superior speech realism and emotional expression. Tables report users preference on the left, and classification accuracy in the confusion matrices. Notably, despite EmoTalk and EMOTE being trained on 10K+ sentences compared to our 40, our model demonstrates enhanced generalization to *in-the-wild* audios (Test 2). A subset of the videos utilized in both user studies can be accessed in the supplementary material.

Table 4. User study comparing E-S2L-S2D and EmoTalk. **Left:** Test 1 and Test 2. **Right:** confusion matrix of Test 3 (in %, “Ours/EmoTalk”, for *Happy, Disgust, Fear, Sad, Angry*).

Test	Criterion	Ours vs. EmoTalk [21]
Test 1	Lip Sync	72.7% / 27.3%
	Emotion	66.6% / 33.4%
Test 2	Lip Sync	76.2% / 23.8%

	HA	DI	FE	SA	AN
HA	100/95.2	0.0/0.0	0.0/0.0	0.0/4.8	0.0/0.0
DI	0.0/14.3	19.0/38.1	4.8/4.8	42.9/9.5	33.3/33.3
FE	0.0/0.0	5.3/14.3	78.9/28.6	10.5/23.8	5.3/33.3
SA	0.0/61.9	9.5/14.3	23.8/4.8	52.3/9.5	14.3/9.5
AN	0.0/0.0	23.8/19.0	4.8/4.8	0.0/14.3	71.4/61.9

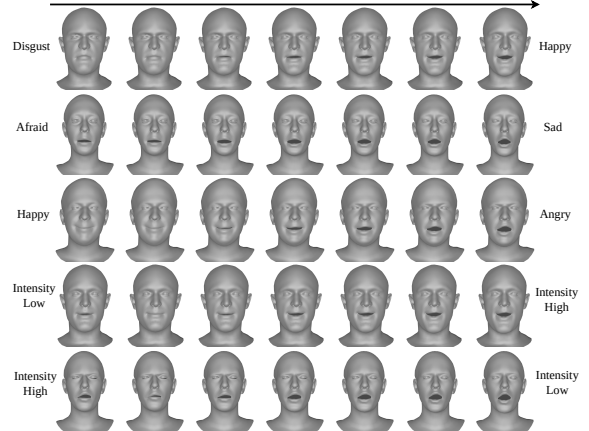


Figure 6. Additional feature of E-S2L+S2D: varying the one-hot labels during inference, we are able to change the emotion or intensity displayed by the talking head.

Table 5. User study comparing E-S2L-S2D and EMOTE. **Left:** Test 1 and Test 2. **Right:** confusion matrix of Test 3 (in %, “Ours/EMOTE”, for *Happy, Disgust, Fear, Sad, Angry*).

Test	Criterion	Ours vs EMOTE [6]
Test 1	Lip Sync	82.4% / 17.6%
	Emotion	55.5% / 45.5%
Test 2	Lip Sync	83.2% / 16.8%

	HA	DI	FE	SA	AN
HA	100/100	0.0/0.0	0.0/0.0	0.0/4.8	0.0/0.0
DI	0.0/0.0	59.1/40.4	0.0/0.0	40.9/0.0	0.0/59.6
FE	0.0/11.0	0.0/0.0	89.8/78.5	0.0/10.5	10.2/0.0
SA	0.0/0.0	0.0/0.0	9.4/9.5	78.4/90.5	12.2/0.0
AN	0.0/0.0	24.3/0.0	0.0/0.0	0.0/0.0	75.7/100

## 5. Conclusions and Limitations

In this paper, we introduced an innovative approach for creating 3D emotional talking heads leveraging only 3D data. Our approach demonstrated strong performance in generating expressive talking heads with high fidelity. Initially, we addressed the scarcity of available datasets by devising a method to combine talking and expressive unpaired 3D faces to generate a new dataset, named EmoVOCA. We defined two versions of this dataset with different sets of emotions. However, our DE-SD framework enables the generation of numerous additional versions by varying the selection of emotional faces used to condition the inexpressive faces in the talking head dataset. Subsequently, we properly customized two state-of-the-art methods, E-Faceformer and E-S2L+S2D. We effectively accomplish the goal by producing more realistic emotional 3D talking heads, if compared to methods that utilize 2D videos as replacement for the lack of 3D data. Our results demonstrate convincing lip syncing and perceivable emotional content.

The proposed approach still faces some potential limitations that can be addressed as future work. First, as shown in the confusion matrices of Tab. 4 and Tab. 5, some emotions resulted more difficult to inject with our solution. In addition, the proposed framework still lacks some realism, like for eyes blinking, head pose, *etc.* How to do so also constitutes an interesting direction to investigate.



## References

- [1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477, 2020. 5
- [2] Giorgos Bouritsas, Sergiy Bokhnyak, Stylianos Ploumpis, Michael Bronstein, and Stefanos Zafeiriou. Neural 3D morphable models: Spiral convolutional networks for 3D shape representation learning and generation. In *IEEE/CVF Int. Conf. on Computer Vision (CVPR)*, pages 7213–7222, 2019. 3, 4
- [3] Che-Jui Chang, Long Zhao, Sen Zhang, and Mubbasir Kapadia. Disentangling audio content and emotion with adaptive instance normalization for expressive facial animation synthesis. *Computer Animation and Virtual Worlds*, 33(3-4):e2076, 2022. 2
- [4] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. Capture, learning, and synthesis of 3D speaking styles. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 4
- [5] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 20311–20322, 2022. 2
- [6] 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*, SA '23, New York, NY, USA, 2023. Association for Computing Machinery. 1, 2, 4, 5, 7, 8
- [7] José Mario De Martino, Léo Pini Magalhães, and Fábio Violaro. Facial animation based on context-dependent visemes. *Computers & Graphics*, 30(6):971–980, 2006. 2
- [8] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: An animator-centric viseme model for expressive lip synchronization. *ACM Trans. on Graphics*, 35(4), jul 2016. 2
- [9] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3D facial animation with transformers. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 18770–18780, 2022. 1, 2, 5, 8
- [10] G.A. Kalberer and L. Van Gool. Face animation based on observed 3D speech dynamics. In *IEEE Conf. on Computer Animation*, pages 20–251, 2001. 2
- [11] 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 Trans. on Graphics*, 36(4), jul 2017. 2
- [12] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 3
- [13] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. 7, 8
- [14] Liying Lu, Tianke Zhang, Yunfei Liu, Xuangeng Chu, and Yu Li. Audio-driven 3D facial animation from in-the-wild videos, 2023. 2
- [15] Federico Nocentini, Thomas Besnier, Claudio Ferrari, Sylvain Arguillere, Stefano Berretti, and Mohamed Daoudi. Beyond fixed topologies: Unregistered training and comprehensive evaluation metrics for 3d talking heads, 2024. 1
- [16] Federico Nocentini, Thomas Besnier, Claudio Ferrari, Sylvain Arguillere, Stefano Berretti, and Mohamed Daoudi. Scantalk: 3d talking heads from unregistered scans, 2024. 1
- [17] Federico Nocentini, Claudio Ferrari, and Stefano Berretti. Learning landmarks motion from speech for speaker-agnostic 3D talking heads generation. In *Int. Conf. on Image Analysis and Processing (ICIAP)*, 2023. 1, 2, 5, 7
- [18] Naima Otberdout, Claudio Ferrari, Mohamed Daoudi, Stefano Berretti, and Alberto Del Bimbo. Sparse to dense dynamic 3D facial expression generation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 20385–20394, 2022. 3, 5
- [19] Wolfgang Paier, Anna Hilsmann, and Peter Eisert. Un-supervised learning of style-aware facial animation from real acting performances. *Graphical Models*, 129:101199, 2023. 1
- [20] Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Self-Talk: A self-supervised commutative training diagram to comprehend 3D talking faces. In *ACM In-*

*ternational Conference on Multimedia, MM '23*, page 5292–5301, 2023. [1](#)

- [21] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. EmoTalk: Speech-driven emotional disentanglement for 3D face animation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20687–20697, October 2023. [1](#), [2](#), [4](#), [5](#), [7](#), [8](#)
- [22] Filippo Principi, Stefano Berretti, Claudio Ferrari, Naima Otberdout, Mohamed Daoudi, and Alberto Del Bimbo. The Florence 4D facial expression dataset. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2023. [1](#), [4](#)
- [23] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3D face animation from speech using cross-modality disentanglement. In *IEEE/CVF Int. Conf. on Computer Vision (CVPR)*, pages 1173–1182, 2021. [1](#), [2](#)
- [24] Yichun Shi, Xiao Yang, Yangyue Wan, and Xiaohui Shen. SemanticStyleGAN: Learning compositional generative priors for controllable image synthesis and editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11244–11254, 2021. [6](#)
- [25] Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. Facediffuser: Speech-driven 3D facial animation synthesis using diffusion. In *ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG'23)*, 2023. [1](#), [2](#), [5](#)
- [26] Kim Sung-Bin, Lee Hyun, Da Hye Hong, Suekyeong Nam, Janghoon Ju, and Tae-Hyun Oh. LaughTalk: Expressive 3D talking head generation with laughter, 2024. [2](#)
- [27] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3D facial animation. In *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 20621–20631, October 2023. [2](#), [5](#), [8](#)
- [28] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12780–12790, 2023. [1](#)