ScanTalk: 3D Talking Heads from Unregistered Scans

Federico Nocentini*[1], Thomas Besnier*[2], Claudio Ferrari[4], Sylvain Arguillere[5], Stefano Berretti[1], and Mohamed Daoudi[2,3]

[1] Media Integration and Communication Center (MICC),
University of Florence, Italy
federico.nocentini@unifi.it, stefano.berretti@unifi.it

[2] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
thomas.besnier@univ-lille.fr

[3] IMT Nord Europe, Institut Mines-Télécom, Centre for Digital Systems
mohamed.daoudi@imt-nord-europe.fr

[4] Department of Architecture and Engineering University of Parma, Italy
claudio.ferrari2@unipr.it

[5] Univ. Lille, CNRS, UMR 8524 Laboratoire Paul Painlevé, Lille, F-59000, France
sylvain.arguillere@univ-lille.fr

Abstract. Speech-driven 3D talking heads generation has emerged as a significant area of interest among researchers, presenting numerous challenges. Existing methods are constrained by animating faces with fixed topologies, wherein point-wise correspondence is established, and the number and order of points remains consistent across all identities the model can animate. In this work, we present **ScanTalk**, a novel framework capable of animating 3D faces in arbitrary topologies including scanned data. Our approach relies on the DiffusionNet architecture to overcome the fixed topology constraint, offering promising avenues for more flexible and realistic 3D animations. By leveraging the power of DiffusionNet, ScanTalk not only adapts to diverse facial structures but also maintains fidelity when dealing with scanned data, thereby enhancing the authenticity and versatility of generated 3D talking heads. Through comprehensive comparisons with state-of-the-art methods, we validate the efficacy of our approach, demonstrating its capacity to generate realistic talking heads comparable to existing techniques. While our primary objective is to develop a generic method free from topological constraints, all state-of-the-art methodologies are bound by such limitations. Code for reproducing our results, and the pre-trained model are available at https://github.com/miccunifi/ScanTalk.

Keywords: 3D Talking Heads · 3D Scans Animation · DiffusionNet

[*] Equal contribution

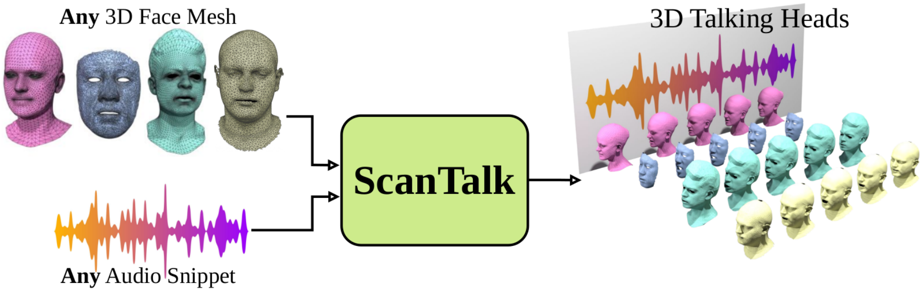


Fig. 1: We present **ScanTalk**, a deep learning architecture to animate **any** 3D face mesh driven by a speech. ScanTalk is robust enough to learn on multiple unrelated datasets with a unique model, whilst allowing us to infer on unregistered face meshes.

1 Introduction

Human face animation is a complex task, widely explored in Computer Vision and Computer Graphics because of the broad range of applications drawn from it, spanning from virtual reality to video game graphics, and more. As 3D face models continue to improve, it becomes more and more relevant to incorporate multi-modal aspects such as speech with audio data. **Speech-driven facial animation** encounters challenges inherent to the fact that finding a cross-modality mapping from audio to geometric data is an ill-posed problem, dealing with the complex geometry of human faces and the limited availability of paired audio-visual 3D data. Another, less discussed aspect that touches a broader range of 3D graphics with meshes is the robustness to changes in the **topology** of the mesh, which refers to the arrangement of the vertices and how they are connected.

Maintaining fidelity and coherence across different topologies is crucial for ensuring that facial animations remain realistic and expressive, regardless of variations in the underlying mesh structure. This challenge becomes particularly pronounced in speech-driven facial animation, where the dynamics of speech-related lip movements, and related face changes necessitate a high degree of adaptability within the mesh topology.

Addressing these challenges requires innovative approaches able to navigate the complexities of facial geometry, while accommodating the nuances of speech-driven animation. In this regard, we present **ScanTalk**, a novel framework capable of animating faces in arbitrary topologies including scanned data. ScanTalk overcomes limitations associated with fixed topologies, offering promising avenues for more flexible and realistic 3D talking heads generation. Indeed, the aforementioned constraint limits the range of applications of current deep learning models for speech-driven motion synthesis. Because of this, deep models are required to train on large scale registered datasets such as FLAME [23] or Multiface [44] for human face data. Building these datasets is a costly procedure, and the trained model is then only usable within the same registered setting. This typically means that any newly acquired data needs to be fitted to the corresponding topology before any animation can be predicted by a deep model. This extra preprocessing step usually prevents online applications, therefore canceling one of the key benefits of deep models, which is their inference speed.

Aiming to address these limitations, in this paper we present a flexible deep learning framework built to generate speech-driven animations of 3D face meshes. In particular, it gathers several key elements:

1. A new robust approach to generate mesh deformation sequences based on DiffusionNet to compute intrinsic descriptors on 3D data;
2. A comprehensive architecture for learning speech-driven animations. Our approach works with meshes with different topologies, while showing competitive performance with respect to state-of-the-art models trained on an individual topology;
3. We show the generalizability of our model to unseen mesh topologies with qualitative examples.

2 Related Work

In literature, there are several methodologies to acquire 3D face meshes, either by extracting the geometry from images or videos [26, 47] or by using a complex set of instruments in a controlled environment [16, 18, 33, 35, 44, 48]. The former offers advantages in terms of easier and more cost-effective data acquisition. However, these methods may sometimes fall short in capturing the complete 3D information from 2D data. With 3D scans, other challenges arise: they come unregistered and may present alterations such as holes and noise. Then, animating these objects directly becomes even more challenging. One can do it by registering the meshes onto a pre-defined topology [16, 22, 23, 28, 44] before animating the registrations with 3D morphable models [14] for example.

Several deep learning models [3–5, 10, 25] proposed ways to learn a latent representation of face scans using robust encoders such as PointNet [6, 31] and Transformers but the resulting mesh is registered, which tends to smooth out some details from the scan geometry. However, this extra registration step may be handled efficiently with recent industrial applications such as MetaHuman from Epic Games. More recently, and closer to our goal, DiffusionNet [37] was combined with neural Jacobian fields [1] in Neural Face Rigging (NFR) [32] to learn a per-triangle deformation field on faces, allowing to transfer an animation from a sequence of unregistered face meshes to another. We stand out from this work by proposing a model that uses audio data to animate a given unregistered face mesh. While DiffusionNet has demonstrated its capability to generalize across varying triangulations in a static setting, our model is the first to exhibit similar properties in a multi-modal 4D setting.

In recent years, numerous models and methodologies have emerged to address the challenge of synchronizing facial animations with speech audio. While significant progress has been made on 2D talking heads [2, 7, 12, 21, 42, 43, 49], only a handful of approaches can be seamlessly extended to 3D data. Procedural techniques have been proposed to animate 3D faces [9, 13, 27, 41, 46], primarily relying on visemes (groups of phonemes) to drive the movements of facial muscles. However, many of these models necessitate re-targeting and struggle to generalize effectively without extensive processing steps. More recent works leverage the increasing availability of data to develop statistical methods, including deep learning strategies [11, 15, 19, 29, 34, 38–40, 45]. Currently, state-of-the-art deep learning approaches are limited to a fixed mesh topology, requiring it to be identical to the one observed during the training phase. Among these models, VOCA [11] pioneers the development of deep models trained on a large-scale dataset of registered meshes. It was outperformed when MeshTalk [34] introduced a larger registered dataset named Multiface [44] along with a more expressive model. Moving forward, FaceFormer [15] utilized the transformer architecture, optimized for temporal data, and leveraged the power of a large-scale pre-trained audio encoder called Wav2vec2 [36]. This strategy saw further development with CodeTalker [45] and SelfTalk [30]. More recently, FaceXHubert [19] and FaceDiffuser [38] utilized an improved audio encoder, Hubert [20], demonstrating superiority to predict cross-modality mappings from audio to face motions. While continuously enhancing performances, these models are bound to a fixed mesh topology, hindering real-world applications and limiting the quantity of usable data for a single model, both for training and inference.

In response to these challenges, we present a novel framework in the landscape of 3D facial animation learning, avoiding the limitations posed by the specific topology of the face mesh to be animated. The proposed model stands out for its capability to animate **any** face mesh, including real-world 3D scans. This novel approach holds the potential to redefine the standards in the field of 3D facial animation, being applicable across diverse datasets and scenarios.

3 Proposed approach: ScanTalk

In this section, we introduce **Scantalk**, a framework for animating 3D face meshes reproducing a spoken sentence contained in an audio file, which does not require the meshes to adhere to any specific topology. With ScanTalk, we push the state-of-the-art in 3D talking heads a step forward by allowing any 3D face, even raw scans, to be animated given a speech. To train ScanTalk, the only requirement is that the training meshes share a common topology *within each sequence*, but the topology may vary from one sequence to another. At inference time, any 3D face mesh in neutral state can be animated.

ScanTalk is an Encoder-Decoder framework, which receives a neutral face mesh and an audio snippet as input, and outputs a sequence of per-vertex deformation fields, whose length depends on that of the audio. By summing each deformation field to the neutral input face, we ultimately obtain the animated sequence. The encoder is composed of two main modules: an audio encoder, which combines a pretrained encoder with a bi-directional LSTM that extracts audio features from the input speech, and a DiffusionNet encoder that computes surface descriptors from the neutral 3D face. These descriptors are then replicated and concatenated to the audio features and the resulting signal is fed to a DiffusionNet decoder that outputs the deformation to be applied to the neutral face. The framework is depicted in Fig. 2. Before providing the technical details in Sec. 3.1 and Sec. 3.2, below we introduce a few general notations.

Let $L = \{(M_i^{gt}, m_i^n, A_i)\}_{i=0}^{N-1}$ denotes the training set comprising N samples, where A_i is an audio containing a spoken sentence, $M_i^{gt} = (m_i^0, \ldots, m_i^{T_i-1}) \in \mathbb{R}^{T_i \times V_i \times 3}$ represents a sequence of 3D faces (same topology) of length T_i synchronized with the spoken sentence in A_i, and $m_i^n \in \mathbb{R}^{V_i \times 3}$ is a 3D neutral face. $V_i = |m_i^n|$ is the number of vertices in the i-th 3D face sequence that needs to be consistent across each mesh in the i-th sequence together with the vertex connectivity, overall determining the topology of the surface. Our objective is to establish a mapping function that correlates an audio input A_i, and a neutral 3D face m_i^n, to the ground-truth sequence M_i^{gt}, expressed as:

$$Scantalk(A_i, m_i^n) \approx M_i^{gt}. \tag{1}$$

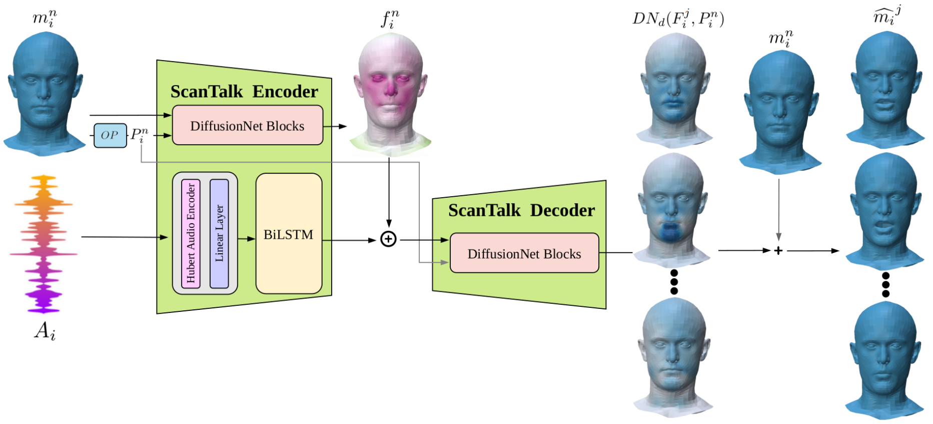


Fig. 2: Architecture of **ScanTalk**. A novel Encoder-Decoder framework designed to dynamically animate any 3D face based on a spoken sentence from an audio file. The Encoder integrates the 3D neutral face m_i^n, per-vertex surface features P_i^n (crucial for DiffusionNet and precomputed by the operators OP), and the audio file A_i, yielding a fusion of per-vertex and audio features. These combined descriptors, alongside P_i^n, are then passed to the Decoder, which mirrors a reversed DiffusionNet encoder structure. The Decoder predicts the deformation of the 3D neutral face, which is then combined with the original 3D neutral face m_i^n to generate the animated sequence.

3.1 ScanTalk Encoder

ScanTalk requires an audio snippet A_i and a 3D face in neutral state m_i^n as inputs. To process the two inputs, ScanTalk comprises two distinct encoders to process the mesh and the audio, respectively, as detailed below.

Face mesh encoder. Several approaches are available for encoding face meshes or point clouds, yet traditional graph convolution-based models, like [17, 24], encounter limitations with varying graph structures, such as changes in mesh resolution. To address this, we employ DiffusionNet [37], a discretization-agnostic encoder that demonstrated to be effective for encoding face meshes as seen in [32]. DiffusionNet integrates multi-layer perceptrons (MLPs), learned diffusion, and spatial gradient features, offering a straightforward yet robust architecture for surface learning tasks. It bypasses the need for complex operations like explicit surface convolutions or pooling hierarchies.

Critical to DiffusionNet's functionality are precomputed features of the face surface, namely the *Cotangent Laplacian*, *Eigenbasis*, *Mass Matrix*, and *Spatial Gradient Matrix*. Together, these operators enhance the architecture's robustness, flexibility, and expressiveness, making DiffusionNet suitable for diverse surface learning tasks. Notably, this architecture accommodates 3D faces of any topology, allowing for variations in the number and order of defining points.
Let $P_i^n = OP(m_i^n)$ represent the precomputed features obtained by applying the above surface closed-form operators OP to the 3D neutral face m_i^n. The DiffusionNet Encoder DN_e, with latent size h, is designed to process a neutral 3D face mesh m_i^n, and requires the precomputed per-vertex features P_i^n to extract a per-vertex descriptors f_i^n, that is:

$$f_i^n = DN_e(m_i^n, P_i^n) \in \mathbb{R}^{V_i \times h}. \tag{2}$$

The per-vertex descriptors f_i^n, is capable of capturing intricate details of each vertex within the neutral 3D face.

Audio encoder. Following [19,38], the speech encoder adopts the architecture of the state-of-the-art pretrained speech model, **HuBERT** [20] that is a self-supervised speech representation learner utilizing an offline clustering step to provide aligned target labels for a BERT-like prediction loss. Using this module, followed by a Linear Layer, we obtain a per-frame audio representation:

$$a_i = SpeechEncoder(A_i) \in \mathbb{R}^{T_i \times (h/2)}. \tag{3}$$

To ensure coherence between the speech representation, following the methodology outlined in [29], we concatenate the *SpeechEncoder* with a Multilayer Bidirectional-LSTM for temporal consistency, in a way such that the speech signal is projected into a **temporal latent representation**:

$$v_i = BiLSTM(a_i) \in \mathbb{R}^{T_i \times h}. \tag{4}$$

3.2 ScanTalk Decoder

We combine per-vertex descriptors f_i^n extracted from the neutral face, with the latent vector v_i extracted from the Bidirectional-LSTM

$$(F_i^j)_k = (f_i^n)_k \oplus v_i^j \in \mathbb{R}^{h*2}, \qquad \forall k = 0, \dots, V_i - 1, \qquad \forall j = 0, \dots, T_i - 1. \quad (5)$$

With this concatenation, we obtain a combined latent $F_i^j \in \mathbb{R}^{V_i \times h*2}$ that embeds both audio-related and geometry-related latents. To decode this sequence of combined latents for deforming the neutral face m_i^n, we employ a DiffusionNet Decoder (which is essentially a reversed Encoder), denoted as DN_d. The decoder module receives F_i^j, and precomputed features P_i^n derived from m_i^n. It predicts the deformation of m_i^n, denoted as $DN_d(F_i^j, P_i^n)$:

$$\widehat{m_i}^j = DN_d(F_i^j, P_i^n) + m_i^n \in \mathbb{R}^{V_i \times 3}. \quad (6)$$

Here $\widehat{m_i}^j$ represents the j-th frame of the predicted sequence. The entire generated sequence is defined by $\widehat{M_i} \in \mathbb{R}^{T_i \times V_i \times 3}$. We opt to utilize ScanTalk for predicting the deformation of the neutral face m_i^n rather than predicting the actual face. This decision aligns with previous works [15,19,30,38,40], and offers advantages in terms of training efficiency and resulting animation. Predicting face deformation makes the learning process easier, by focusing solely on speech-related motion, as opposed to incorporating the problem of predicting the entire face reconstruction. Essentially, the decoder learns to predict a per-vertex displacement field from a time-dependent per-vertex descriptors field.

3.3 ScanTalk Training

ScanTalk generates deformations of a neutral face; hence, a predicted sequence maintains a consistent topology across all frames. This property arises from the definition of the DiffusionNet Decoder, which necessitates knowledge of precomputed features P_i^n, and the number of points in the 3D neutral face targeted for animation. For this reason, during a supervised training protocol, the ground-truth meshes within each sequence must adhere to a common topology. Despite the apparent specificity of this requirement, it proves to be non-prohibitive in practice. ScanTalk animates a neutral face in response to an audio input, producing a sequence of 3D faces sharing the same topology of the neutral face that is animated, which can be any topology (even different from those seen during training). This alignment between the training strategy and the desired inference outcome underscores the efficiency of ScanTalk, which is not restricted to the topology of the training faces.

Notably, the model predicts per-vertex displacements of the neutral ground-truth mesh and the ground truth sequence is registered in a supervised setting. Consequently, the predicted sequence and the ground-truth sequence are aligned on the same topology but can vary from one sequence to another. Thus, during training, we minimize the average vertex-wise Mean Squared Error (MSE) over a sequence of length T_i between the ground truth M_i^{gt} and the model prediction $\widehat{M_i}$. which is defined as:

$$\mathcal{L}_{MSE} = \frac{1}{T_i - 1} \sum_{j=0}^{T_i-1} \frac{1}{V_i - 1} \sum_{k=0}^{V_i-1} \left\| (m_i^j)_k - (\widehat{m_i}^j)_k \right\|_2^2 . \quad (7)$$

4 Experiments

In the following, we first introduce the datasets and the metrics used for evaluation, respectively in Sec. 4.1 and Sec. 4.2. Then, we report quantitative and qualitative results in comparison with state-of-the-art methods in Sec. 4.3 and Sec. 4.4. In Sec. 4.5 we present several ablation studies over the framework architecture. Finally, in Sec. 4.6, we report the results of an user study designed to compare our approach with state-of-the-art methods.

4.1 Datasets

For training and quantitative evaluations, we rely on three classical datasets in 3D speech-driven motion synthesis: VOCAset [11], BIWI [16] and Multiface [44].
VOCAset gathers mesh sequences of 12 actors performing 40 speeches, captured at 60fps. Each sequence lasts for around 3 to 5 seconds, and each mesh is registered to the FLAME [23] topology with 5,023 vertices and 9,976 faces.
BIWI comprises 14 subjects articulating 40 sentences each, sampled at 25fps. Each sentence lasts around 5 seconds, and the meshes are registered to a fixed topology. Due to GPU limitations, we used a downsampled version of this dataset, called BIWI$_6$, which has a fixed topology, with 3,895 vertices and 7,539 faces.
Multiface includes 13 identities executing up to 50 speeches of around 4 seconds each, sampled at 30fps. The original multiface meshes have a fixed topology with 5,471 vertices and 10,837 faces.

For training purposes, the meshes in both **Multiface** and **BIWI$_6$** have been scaled and aligned with the meshes in the **VOCAset** dataset. The data splits can be found in the supplementary material. Since ScanTalk is the first topology-independent 3D talking head generator, for the sake of a comprehensive comparison with the state-of-the-art, we trained 4 different models: one is trained using all the three datasets together (multi-dataset), while the other three are trained on each dataset separately (single-dataset).

4.2 Evaluation Metrics

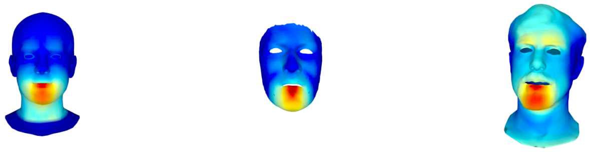
To evaluate the quality of the generated faces, we employ three standard metrics from previous works [30,34,38,45], namely: **Lip vertex error** (LVE) $\times 10^{-5}(mm)$. Assesses the lip deviation in a sequence by comparing it to the ground truth. It is obtained by computing the maximum L2 error for all lip vertices in each frame, averaged across all frames.

Mean vertex error (MVE) $\times 10^{-3}(mm)$. Describes the quality of the overall face reconstruction by computing the maximal vertex L2 error for each frame and taking the mean across all frames.
Upper-face dynamic deviation (FDD) $\times 10^{-7}(mm)$. Compares the standard deviation of upper vertices between the generated sequence and ground-truth.

4.3 Quantitative Results

Although ScanTalk is the first deep learning model to animate unregistered 3D face meshes, comparisons with the state-of-the-art is possible when inferring our model on registered data. In Tab. 1, we present the performance evaluation of ScanTalk training on both a single-dataset and multi-datasets, juxtaposed against several state-of-the-art methods trained solely on a single dataset. While the augmentation of training data enhances the effectiveness of our model, addressing disparities in data composition becomes important for animating unseen faces. Particularly, the varied geometries and head dynamics inherent in BIWI$_6$, VOCA, and Multiface meshes pose significant challenges, potentially limiting performance compared to a model trained exclusively on a single dataset.

Table 1: ScanTalk performance, in both single-dataset (s-d) and multi-dataset (m-d) scenarios, in comparison with state-of-the-art methods. The heatmaps show the differences between the first frame and subsequent frames within sequences among the VOCAset, BIWI$_6$, and Multiface datasets. Notably, in the VOCAset, primarily the lips display movement, whereas in BIWI$_6$ and Multiface datasets, substantial head and upper face movements are observed. The color gradient on the face meshes corresponds to the average per-vertex L_2 norm of the differences, where blue hues indicate lower values, and red hues indicate higher values.



	VOCAset			**BIWI$_6$**			**Multiface**		
	LVE ↓	MVE ↓	FDD ↓	LVE ↓	MVE ↓	FDD ↓	LVE ↓	MVE ↓	FDD ↓
VOCA	6.993	0.983	2.662	5.743	2.591	41.482	4.923	2.761	55.781
FaceFormer	6.123	0.935	2.163	4.085	2.163	37.091	2.451	**1.453**	**20.239**
SelfTalk	5.618	0.918	2.321	**3.628**	2.062	35.470	**2.281**	1.901	37.434
CodeTalker	3.549	0.888	2.258	5.190	2.641	**20.599**	4.091	2.382	47.905
FaceDiffuser	4.350	0.901	2.437	4.022	2.128	39.604	3.555	2.388	29.157
ScanTalk m-d	6.375	0.987	**2.101**	4.044	**2.057**	40.051	2.435	1.678	32.202
ScanTalk s-d	**3.012**	**0.861**	2.400	4.651	2.148	36.034	2.653	1.871	64.451

This observation becomes evident when assessing Scantalk's performance on VOCAset, where a substantial improvement is observed when training on a single-dataset. Conversely, in the case of BIWI$_6$ and Multiface single-dataset training, a decline in results is noted compared to the multi-dataset version. This disparity is attributed to the presence of speech-related head and upper face movements within the ground truth talking head sequences of BIWI$_6$ and Multiface. Consequently, the multi-dataset training model demonstrates superior performance as it encounters a wider array of sequences featuring head and upper face movements. In contrast, VOCAset lacks such movements, with the 3D faces remaining static, while only the mouth moves, thereby enhancing performance for ScanTalk trained solely on VOCAset. This obviates the need to learn other movements of the head or upper face.

The distinctions between VOCAset and the other two datasets are further underscored by the FDD metric, indicating that 3D faces in VOCAset remain stationary, with minimal activity in the upper facial area and head. Across all tested models, generating sequences with lower FDD is notably more achievable on VOCAset compared to the other datasets, thus emphasizing the disparities between VOCAset and the others.

Nevertheless, the results presented in Tab. 1 demonstrate that ScanTalk produces outcomes comparable to state-of-the-art methods, whether undergoing single-dataset or multi-dataset training. Moreover, ScanTalk emerges as the first model capable of training in multi-dataset settings, demonstrating the ability to animate a 3D face regardless of its topology. An interesting characteristic of ScanTalk emerges when analyzing the GPU memory usage relative to the vertex count in the 3D facial model requiring animation. Fig. 3 illustrates a linear increase in GPU usage correlated with the number of vertices.

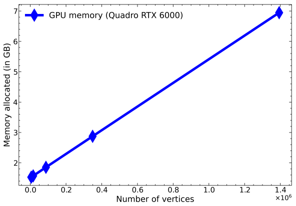


Fig. 3: ScanTalk GPU memory usage with respect to the mesh resolution.

4.4 Qualitative Results

In the domain of 3D speech-driven talking heads generation, qualitative evaluations hold as much significance as quantitative assessments. To evaluate the quantitative findings presented in Sec. 4.3, we introduce a straightforward test aimed at discerning disparities in head and upper face movements across the three datasets: for each dataset, we compute the average L_2 norm of the differences between the initial frame and subsequent frames within each sequence. This enables the quantification of both head dynamics and upper facial movements across mesh sequences. In the heatmaps of Tab. 1, we present the per-frame average difference for each dataset. Our analysis reveals that in the VOCAset, only the lower part of the face exhibits discernible movements, whereas in BIWI$_6$ and Multiface, the entire head or face moves.

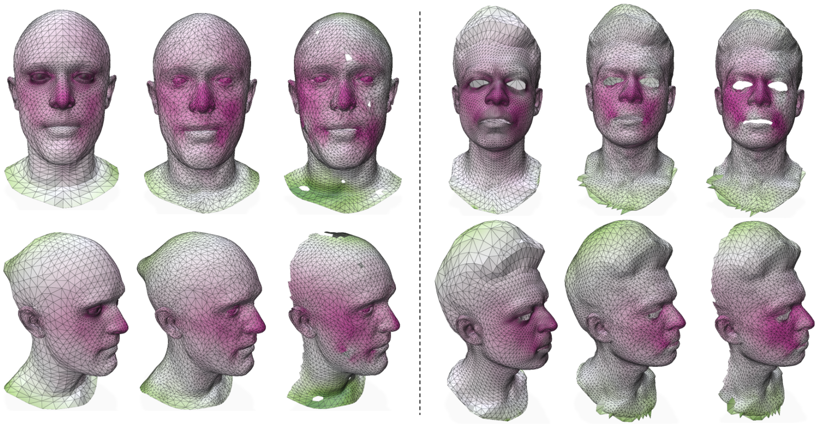


Fig. 4: Relative norm of the per-vertex descriptors f_i^n in Eq. (2) extracted by DN_e displayed as a heatmap on a mesh from VOCAset (left), and a mesh from Multiface (right). For each mesh, we show the norm on the original topology, on a remeshed version, and on a further degraded mesh obtained by removing the back of the head and creating random holes. Here, pinker hues indicate lower values, and greener hues indicate higher values.

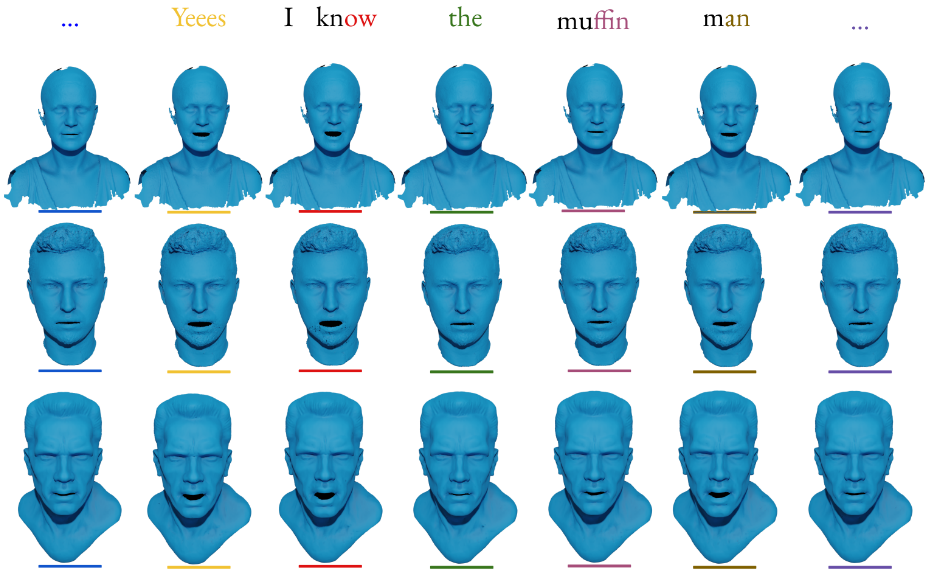


Fig. 5: ScanTalk inference on 3D faces. The meshes have been rigidly aligned with the training data. The first row is an animation of a raw 3D scan (an hole for the mouth has been created), while the others two are animation of meshes in an arbitrary topology.

In Fig. 4, we present a heatmap visualization representing the norm of geometric descriptors f_i^n computed by the encoder DN_e in Eq. (2) for various alterations of the same identity mesh. Our findings indicate that the ScanTalk encoder effectively extracts descriptors from input faces regardless of facial topology. Notably, despite differences in facial topology among the same identity, the extracted descriptors exhibit remarkable similarity.

In Fig. 5, 3D faces in different topology, including a scan, animated using ScanTalk are presented. Our method demonstrates good generalization capabilities, successfully animating diverse 3D faces. Notably, enhanced animation quality is observed when a mouth aperture is present; however, our model performs well even in its absence. Nonetheless, without a mouth aperture, the model struggles to generate the corresponding mouth opening, despite accurate lip movement synchronization.

4.5 Ablations Studies

To investigate the impacts of various components within our architecture, we conducted extensive experiments across various configurations, encompassing both single-dataset and multi-dataset training paradigms. The results reported in Tab. 2 and Tab. 3 are computed by modifying ScanTalk's modules and training losses as described in Sec. 3.

Audio encoder. In learning based speech-driven animation, some previous work used either Wav2vec2 [36] or Hubert [20], the latter being substantially better for this task according to [19,38]. For our purpose, we also tested WavLM [8] but obtained poorer efficiency. The results on both the single-dataset and multi-dataset training are displayed in Tab. 2a and Tab. 3 top. We can see that, in both cases, the usage of the Hubert Encoder [19] for audio features extraction leads to better results.

Temporal consistency. In our investigation, as proven by [19,29,40], we found that supplementing pre-trained audio encoders with additional temporal consistency mechanisms such as Bidirectional-LSTM, Bidirectional-GRU, or an autoregressive Transformer Decoder (TD) significantly enhances model performance, as illustrated in Tab. 2b and in Tab. 3 middle. The BiLSTM model architecture is detailed in Sec. 3, while the BiGRU model mirrors the BiLSTM architecture but substitutes BiLSTM with BiGRU. Details about the Transformer Decoder can be found in the supplementary material. From the findings presented in Tab. 2b and in Tab. 3 middle, we observe that employing a multilayer Bidirectional-LSTM in the audio stream processing of ScanTalk yields the most favorable performance across both single-dataset training and multi-dataset training scenarios.

Loss function. Previous research has extensively explored optimal objective functions for enhancing and refining the learning process. Inspired by [29,30,40], we experimented with Mean Square Error (L_{MSE}), Masked (L_{mask}), and Velocity (L_{vel}) Loss. The former is a simple L_2 loss, the second employs Mean Square Error focused on lip vertices, while the latter aims to minimize differences between consecutive frames. While these loss functions have demonstrated efficacy in single-dataset training, as evidenced by [29,30,40], our findings, detailed in Tab. 3 bottom, indicate that employing a straightforward L_2, on multi-dataset training, enhances the model's capacity to generate realistic talking heads. We attribute this to significant geometric variations observed across different datasets.

Table 2: ScanTalk (ST) single-dataset ablation studies. Results obtained with a model trained just on VOCAset.

(a) Audio Encoder Ablation.

	LVE ↓	MVE ↓	FDD ↓
ST w WavLM	3.674	0.937	2.413
ST w Wav2Vec2	3.309	**0.860**	**2.244**
ScanTalk	**3.012**	0.861	2.400

(b) Temporal consistency ablation.

	LVE ↓	MVE ↓	FDD ↓
ST w/o	3.361	0.870	2.365
ST w TD	3.291	0.859	2.406
ST w BiGRU	3.036	**0.835**	**2.358**
ScanTalk	**3.012**	0.861	2.400

Table 3: ScanTalk (ST) multi-dataset ablation studies. **Top**: ST with different audio encoders. **Middle**: ST using different audio stream processing. **Bottom**: ST with different Loss function. The last row is the proposed ScanTalk

	VOCAset			**BIWI$_6$**			**Multiface**		
	LVE ↓	MVE ↓	FDD ↓	LVE ↓	MVE ↓	FDD ↓	LVE ↓	MVE ↓	FDD ↓
ST w Wav2Vec2	7.127	1.241	2.349	4.439	2.276	41.374	3.198	2.329	33.671
ST w WavLM	6.773	1.004	2.220	4.333	2.162	**31.922**	2.907	2.004	**8.500**
ST w TD	6.512	0.995	1.858	4.827	2.168	36.789	2.136	1.725	35.044
ST w BiGRU	6.584	0.994	2.054	4.421	2.103	41.456	2.534	1.987	34.098
ST+$L_{mask} + L_{vel}$	7.031	1.092	1.473	4.755	2.227	39.458	2.547	1.898	15.973
ST+L_{mask}	7.451	1.084	**0.859**	4.748	2.199	36.939	2.888	1.891	32.501
ST+L_{vel}	6.740	0.998	1.899	4.509	2.180	33.614	**2.103**	1.775	31.599
ScanTalk	**6.375**	**0.987**	2.101	**4.044**	**2.057**	40.051	2.435	**1.678**	32.202

4.6 User study

To further evaluate our solution, we performed a study where human feedback is involved. This evaluation is conducted with two user-based studies involving 25 participants; *(i)* For the first study, in alignment with prior research [15,19,30,34, 38,45], we designed an A/B test to compare ScanTalk with other state-of-the-art models within a registered setting on both lip-syncing and naturalness criteria (Test 1); *(ii)* In the second test, we assessed the credibility of scan animations by asking participants to rate the animation quality of ten scans sourced from the COMA dataset [33], using a scale ranging from 1 to 10 (Test 2).

The outcomes of Test 1 are depicted in the table on the left of Fig. 6. Notably, sequences produced using ScanTalk demonstrate levels of both naturalness and lip-syncing that are comparable to state-of-the-art methods. Specifically, our approach is preferred when compared with FaceFormer, CodeTalker, and FaceDiffuser on VOCAset. However, SelfTalk and the ground truth are preferred over ScanTalk. Nevertheless, the percentage of users favoring ScanTalk remains non-negligible, underscoring the efficacy of our approach in generating 3D talking heads with good realism and lip-syncing fidelity.

Results of Test 2 are reported on the right of Fig. 6. These ratings are closely linked to both the scan quality and the fidelity of the mouth representation. Despite the scan quality not being optimal, as depicted in Fig. 6, our results indicate that we can achieve a good level of realism in the animated scans.

Comparison	Naturalness (%)	Lip-sync (%)
ST vs. FaceFormer	82.67	80.67
ST vs. CodeTalker	57.33	56.00
ST vs. FaceDiffuser	90.67	94.00
ST vs. SelfTalk	44.00	43.33
ST vs. GT	44.00	42.00

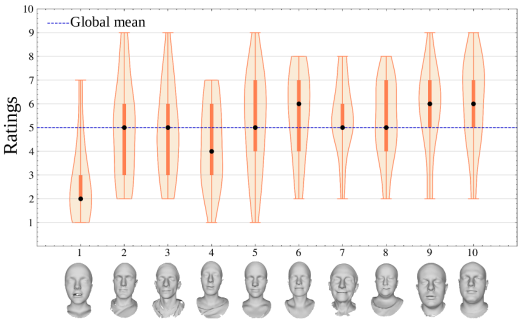


Fig. 6: The table on the left shows results of Test 1, ScanTalk (ST) vs. other methods and the ground truth (GT) on samples from the test set of VOCAset. The values denote the percentages of users who favored ScanTalk animations over others. The chart on the right shows results of Test 2 reported as a violin plot for the animation of scans from COMA [33]. The subject median rating is displayed as a black dot on each violin.

5 Conclusions

This paper introduces ScanTalk, a novel framework for speech-driven 3D facial animation. Unlike existing methods, ScanTalk possesses the unique capability to animate any 3D face, regardless of its topology, even if it differs from the ones on which it was trained. ScanTalk extends the applicability of deep speech-driven 3D facial animations by addressing the challenges of topology robustness. Additionally, our model demonstrates comparable quantitative and qualitative results with other state-of-the-art methods across three distinct datasets.

Acknowledgments This work is supported by the ANR project Human4D ANR-19-CE23-0020 and by the IRP CNRS project GeoGen3DHuman. This work was also partially supported by "Partenariato FAIR (Future Artificial Intelligence Research) - PE00000013, CUP J33C22002830006" funded by NextGenerationEU through the italian MUR within NRRP, project DL-MIG. This work was also partially funded by the ministerial decree n.352 of the 9th April 2022 NextGenerationEU through the italian MUR within NRRP. This work was also partially supported by Fédération de Recherche Mathématique des Hauts-de-France (FMHF, FR2037 du CNRS).

References

1. Aigerman, N., Gupta, K., Kim, V.G., Chaudhuri, S., Saito, J., Groueix, T.: Neural jacobian fields: Learning intrinsic mappings of arbitrary meshes. ACM Trans. Graph. **41**(4) (jul 2022). `https://doi.org/10.1145/3528223.3530141`, `https://doi.org/10.1145/3528223.3530141`
2. Alghamdi, M.M., Wang, H., Bulpitt, A.J., Hogg, D.C.: Talking head from speech audio using a pre-trained image generator. In: Proceedings of the 30th ACM International Conference on Multimedia (2022)
3. Bahri, M., O' Sullivan, E., Gong, S., Liu, F., Liu, X., Bronstein, M.M., Zafeiriou, S.: Shape my face: Registering 3d face scans by surface-to-surface translation. International Journal of Computer Vision (IJCV) (Sep 2021)
4. Besnier, T., Arguillère, S., Pierson, E., Daoudi, M.: Toward Mesh-Invariant 3D Generative Deep Learning with Geometric Measures. Computers and Graphics (2023). `https://doi.org/10.1016/j.cag.2023.06.027`, `https://hal.science/hal-04143649`
5. Chandran, P., Zoss, G., Gross, M., Gotardo, P., Bradley, D.: Shape transformers: Topology-independent 3d shape models using transformers. In: Computer Graphics Forum. vol. 41, pp. 195–207. Wiley Online Library (2022)
6. Charles, R.Q., Su, H., Kaichun, M., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). p. 77–85. IEEE (Jul 2017)
7. Chen, L., Cui, G., Liu, C., Li, Z., Kou, Z., Xu, Y., Xu, C.: Talking-head generation with rhythmic head motion. In: European Conference on Computer Vision (2020), `https://api.semanticscholar.org/CorpusID:220633152`
8. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Zeng, M., Wei, F.: Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing **16**, 1505–1518 (2021), `https://api.semanticscholar.org/CorpusID:239885872`
9. Cosi, P., Caldognetto, E., Perin, G., Zmarich, C.: Labial coarticulation modeling for realistic facial animation. In: Proceedings. Fourth IEEE International Conference on Multimodal Interfaces. pp. 505–510 (2002). `https://doi.org/10.1109/ICMI.2002.1167047`
10. Croquet, B., Christiaens, D., Weinberg, S.M., Bronstein, M., Vandermeulen, D., Claes, P.: Unsupervised diffeomorphic surface registration and non-linear modelling. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). p. 118–128. Springer (2021)

11. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.: Capture, learning, and synthesis of 3D speaking styles. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 10101–10111 (2019)
12. Das, D., Biswas, S., Sinha, S., Bhowmick, B.: Speech-driven facial animation using cascaded gans for learning of motion and texture. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 408–424. Springer International Publishing, Cham (2020)
13. Edwards, P., Landreth, C., Fiume, E., Singh, K.: Jali: an animator-centric viseme model for expressive lip synchronization. ACM Trans. Graph. **35**(4) (jul 2016). `https://doi.org/10.1145/2897824.2925984`, `https://doi.org/10.1145/2897824.2925984`
14. Egger, B., Smith, W.A.P., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., Theobalt, C., Blanz, V., Vetter, T.: 3D Morphable Face Models - Past, Present and Future. ACM Transactions on Graphics **39**(5), 157:1–38 (Aug 2020). `https://doi.org/10.1145/3395208`, `https://inria.hal.science/hal-02280281`
15. Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T.: Faceformer: Speech-driven 3d facial animation with transformers. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 18749–18758. IEEE, New Orleans, LA, USA (Jun 2022). `https://doi.org/10.1109/CVPR52688.2022.01821`, `https://ieeexplore.ieee.org/document/9878591/`
16. Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., Van Gool, L.: A 3-d audio-visual corpus of affective communication. IEEE Transactions on Multimedia **12**(6), 591–598 (2010). `https://doi.org/10.1109/TMM.2010.2052239`
17. Gong, S., Chen, L., Bronstein, M., Zafeiriou, S.: Spiralnet++: A fast and highly efficient mesh convolution operator. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019)
18. Gupta, S., Castleman, K.R., Markey, M.K., Bovik, A.C.: Texas 3d face recognition database. In: 2010 IEEE Southwest Symposium on Image Analysis & Interpretation (SSIAI). pp. 97–100 (2010). `https://doi.org/10.1109/SSIAI.2010.5483908`
19. Haque, K.I., Yumak, Z.: 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). ACM, New York, NY, USA (2023). `https://doi.org/10.1145/3577190.3614157`, `https://doi.org/10.1145/3577190.3614157`
20. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans. Audio, Speech and Lang. Proc. **29**, 3451–3460 (oct 2021). `https://doi.org/10.1109/TASLP.2021.3122291`, `https://doi.org/10.1109/TASLP.2021.3122291`
21. Ji, X., Zhou, H., Wang, K., Wu, Q., Wu, W., Xu, F., Cao, X.: Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In: ACM SIGGRAPH 2022 Conference Proceedings. SIGGRAPH '22 (2022). `https://doi.org/10.1145/3528233.3530745`, `https://doi.org/10.1145/3528233.3530745`
22. Li, J., Kuang, Z., Zhao, Y., He, M., Bladin, K., Li, H.: Dynamic facial asset and rig generation from a single scan. ACM Trans. Graph. **39**(6) (nov 2020). `https://doi.org/10.1145/3414685.3417817`, `https://doi.org/10.1145/3414685.3417817`
23. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: 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), `https://doi.org/10.1145/3130800.3130813`

24. Lim, I., Dielen, A., Campen, M., Kobbelt, L.: A simple approach to intrinsic correspondence learning on unstructured 3d meshes. p. 349–362. Springer-Verlag, Berlin, Heidelberg (2019). https://doi.org/10.1007/978-3-030-11015-4_26, https://doi.org/10.1007/978-3-030-11015-4_26
25. Liu, F., Tran, L., Liu, X.: 3d face modeling from diverse raw scan data. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9407–9417. IEEE Computer Society, Los Alamitos, CA, USA (nov 2019). https://doi.org/10.1109/ICCV.2019.00950
26. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M., Lee, J., Chang, W.T., Hua, W., Georg, M., Grundmann, M.: Mediapipe: A framework for perceiving and processing reality. In: Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019 (2019), https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf
27. Massaro, D., Cohen, M., Tabain, M., Beskow, J., Clark, R.: Animated speech: Research progress and applications. Audiovisual Speech Processing (01 2001). https://doi.org/10.1017/cbo9780511843891.014
28. Muralikrishnan, S., Huang, C.H.P., Ceylan, D., Mitra, N.J.: Bliss: Bootstrapped linear shape space (2023)
29. Nocentini, F., Ferrari, C., Berretti, S.: Learning landmarks motion from speech for speaker-agnostic 3d talking heads generation. In: Foresti, G.L., Fusiello, A., Hancock, E. (eds.) Image Analysis and Processing – ICIAP 2023. pp. 340–351. Springer Nature Switzerland, Cham (2023)
30. Peng, Z., Luo, Y., Shi, Y., Xu, H., Zhu, X., Liu, H., He, J., Fan, Z.: Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In: Proceedings of the 31st ACM International Conference on Multimedia. p. 5292–5301 (2023). https://doi.org/10.1145/3581783.3611734
31. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems (NeurIPS) **30** (2017)
32. Qin, D., Saito, J., Aigerman, N., Thibault, G., Komura, T.: Neural face rigging for animating and retargeting facial meshes in the wild. In: SIGGRAPH 2023 Conference Papers (2023)
33. Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3D faces using convolutional mesh autoencoders. In: European Conference on Computer Vision (ECCV). pp. 725–741 (2018), http://coma.is.tue.mpg.de/
34. Richard, A., Zollhöfer, M., Wen, Y., de la Torre, F., Sheikh, Y.: Meshtalk: 3d face animation from speech using cross-modality disentanglement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1173–1182 (October 2021)
35. Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L.: Bosphorus database for 3d face analysis. In: Schouten, B., Juul, N.C., Drygajlo, A., Tistarelli, M. (eds.) Biometrics and Identity Management. pp. 47–56. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)
36. Schneider, S., Baevski, A., Collobert, R., Auli, M.: wav2vec: Unsupervised pre-training for speech recognition. In: Interspeech 2019. p. 3465–3469. ISCA (Sep 2019). https://doi.org/10.21437/Interspeech.2019-1873, https://www.isca-speech.org/archive/interspeech_2019/schneider19_interspeech.html
37. Sharp, N., Attaiki, S., Crane, K., Ovsjanikov, M.: Diffusionnet: Discretization agnostic learning on surfaces. ACM Trans. Graph. **01**(1) (2022)

38. Stan, S., Haque, K.I., Yumak, Z.: Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In: ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG '23), November 15–17, 2023, Rennes, France. ACM, New York, NY, USA (2023). https://doi.org/10.1145/3623264.3624447, https://doi.org/10.1145/3623264.3624447
39. Thambiraja, B., Aliakbarian, S., Cosker, D., Thies, J.: 3diface: Diffusion-based speech-driven 3d facial animation and editing (2023)
40. Thambiraja, B., Habibie, I., Aliakbarian, S., Cosker, D., Theobalt, C., Thies, J.: Imitator: Personalized speech-driven 3d facial animation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20621–20631 (October 2023)
41. Wang, A., Emmi, M., Faloutsos, P.: Assembling an expressive facial animation system. In: Proceedings of the 2007 ACM SIGGRAPH Symposium on Video Games. p. 21–26. Sandbox '07, Association for Computing Machinery, New York, NY, USA (2007). https://doi.org/10.1145/1274940.1274947, https://doi.org/10.1145/1274940.1274947
42. Wang, J., Zhao, Y., Liu, L., Xu, T.S., Li, Q., Li, S.: Emotional talking head generation based on memory-sharing and attention-augmented networks. ArXiv **abs/2306.03594** (2023), https://api.semanticscholar.org/CorpusID:259089214
43. Wang, S., Li, L., Ding, Y., Yu, X.: One-shot talking face generation from single-speaker audio-visual correlation learning. In: AAAI 2022 (2022)
44. Wuu, C.h., Zheng, N., Ardisson, S., Bali, R., Belko, D., Brockmeyer, E., Evans, L., Godisart, T., Ha, H., Huang, X., Hypes, A., Koska, T., Krenn, S., Lombardi, S., Luo, X., McPhail, K., Millerschoen, L., Perdoch, M., Pitts, M., Richard, A., Saragih, J., Saragih, J., Shiratori, T., Simon, T., Stewart, M., Trimble, A., Weng, X., Whitewolf, D., Wu, C., Yu, S.I., Sheikh, Y.: Multiface: A dataset for neural face rendering. In: arXiv (2022). https://doi.org/10.48550/ARXIV.2207.11243, https://arxiv.org/abs/2207.11243
45. Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J., Wong, T.T.: Codetalker: Speech-driven 3d facial animation with discrete motion prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12780–12790 (2023)
46. Xu, Y., Feng, A.W., Marsella, S., Shapiro, A.: A practical and configurable lip sync method for games. In: Proceedings of Motion on Games. p. 131–140. MIG '13, Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2522628.2522904, https://doi.org/10.1145/2522628.2522904
47. Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., Black, M.J.: Generating holistic 3d human motion from speech. In: CVPR (2023)
48. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.: A 3d facial expression database for facial behavior research. In: 7th International Conference on Automatic Face and Gesture Recognition (FGR06). pp. 211–216 (2006). https://doi.org/10.1109/FGR.2006.6
49. Zhong, W., Fang, C., Cai, Y., Wei, P., Zhao, G., Lin, L., Li, G.: Identity-preserving talking face generation with landmark and appearance priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9729–9738 (June 2023)