ScanTalk: 3D Talking Heads from Unregistered Scans

Accepted for **ECCV 2024** main conference

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

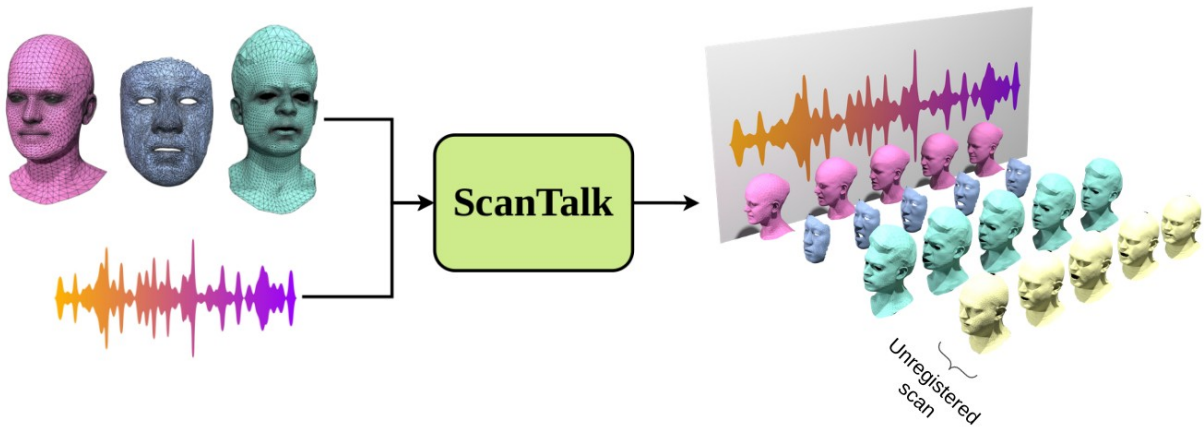
federico.nocentini@unifi.it, stefano.berretti@unifi.it

[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



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. However, 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 to ensure realistic and expressive facial animations, 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.
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:

- A new robust approach to generate mesh deformation sequences based on DiffusionNet to compute intrinsic descriptors on 3D data;
- A comprehensive architecture for learning speech-driven animations. We show the relevance of 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;
- We show the generalizability of our model to unseen mesh topologies with qualitative examples.

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.

ScanTalk is an Encoder-Decoder framework taking a neutral 3D face mesh and an audio snippet as input, outputting a sequence of per-vertex deformation fields to animate the face. The encoder is made of an audio encoder and a geometry encoder for surface descriptors. These descriptors are concatenated with audio features and fed into the DiffusionNet decoder, which predicts the deformation applied to the neutral face to generate the animated sequence. ScanTalk represents a significant advancement in 3D speech-driven talking heads generation, overcoming the limitations of fixed topologies in existing models. By enabling animation of any 3D face mesh, including raw scans, ScanTalk holds potential for diverse applications and sets a new standard in 3D facial animation. The framework's adaptability to various topologies and its competitive performance compared to state-of-the-art methods highlight its contribution to the field of computer vision and graphics.