UNIVERSITÀ DEGLI STUDI FIRENZE

Speech-Driven 3D Talking Heads using Facial Landmarks

Authors: Federico Nocentini, Claudio Ferrari, Stefano Berretti
Department of Information Engineering
Laboratory: MICC (Media Integration and Communication Center)

PhD program in Smart Computing

Abstract

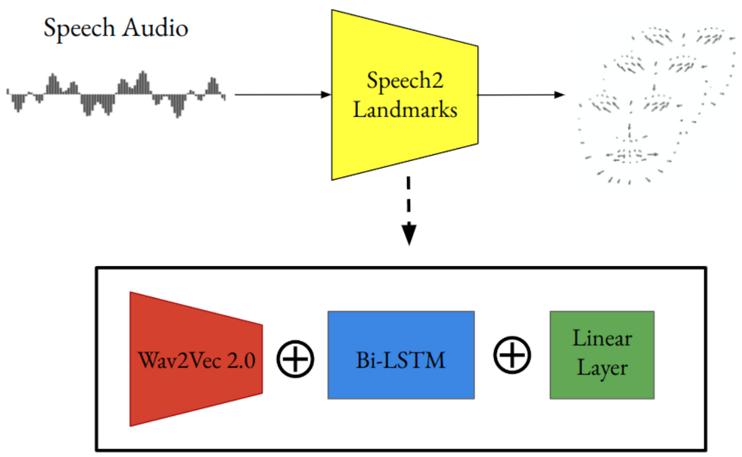
This paper presents a novel approach for generating 3D talking heads from raw audio inputs. Our method grounds on the idea that speech related movements can be comprehensively and efficiently described by the motion of a few control points located on the movable parts of the face *i.e.* landmarks. The underlying musculoskeletal structure then allows us to use the landmarks motion to model geometrical deformations of the whole face. The proposed method employs two distinct models to this aim: the first one learns to generate the motion of a sparse set of landmark from the given audio. The second model expands such landmarks motion to a dense motion field, which is utilized to animate a given 3D mesh in neutral state. Additionally, we introduce a novel loss function, named Cosine Loss, which minimizes the angle between the generated motion vectors and the ground truth ones.

Introduction

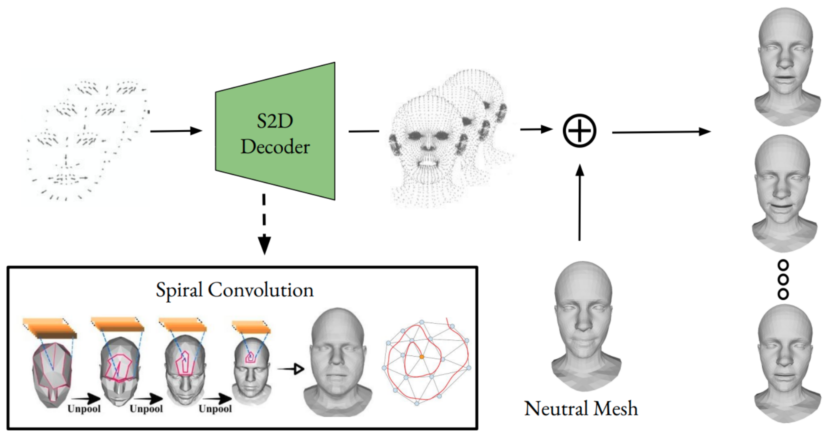
Speech-driven 3D talking heads generation is a rapidly growing field of research and development that has garnered significant interest in recent years. This technology involves generating realistic 3D digital avatars that can accurately replicate human speech and facial expressions in real-time. This innovation has far-reaching implications for a wide range of applications, including virtual assistants, video games, education, and entertainment.

Proposed Approach

We introduce a novel approach for generating 3D talking heads that decomposes the problem into two distinct sub-problems, each tackled by a separate model. The first model (**S2L**) tracks the movements of scattered landmarks in response to the speech. Specifically, it takes an audio signal as input, from which it generates a frame-by-frame motion of a set of landmarks. The motion is modeled as displacement relative to a neutral configuration of 3D landmarks.



The second model (**S2D**) takes the resulting displacement of scattered landmarks and densifies them to create a dense motion field. Using the latter, the model then animates a 3D face mesh by adding the motion field to the 3D face vertices.



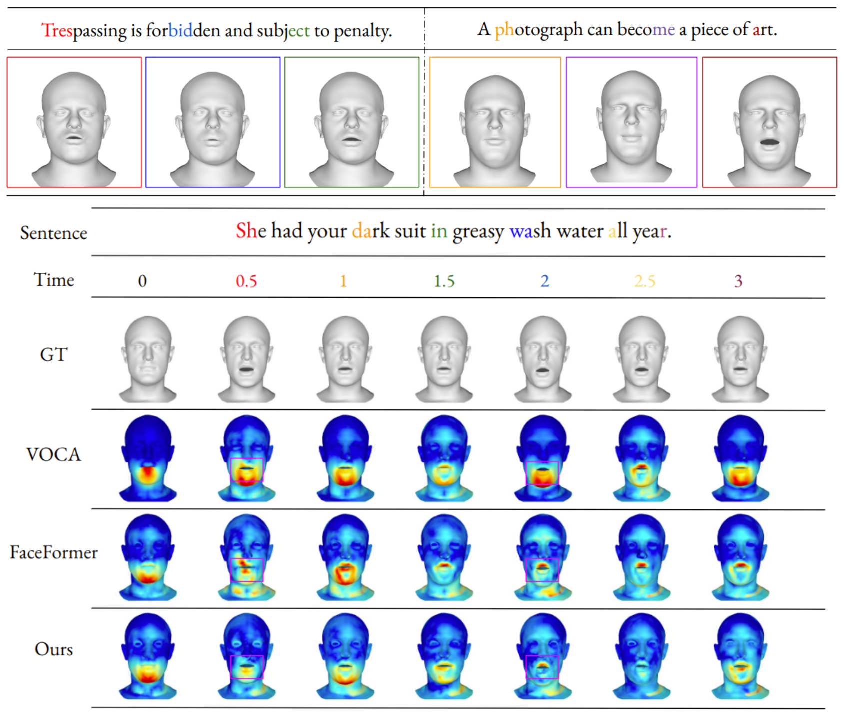
By addressing each sub-problem independently, we aim to improve the overall performance and efficiency of our approach for generating high-quality 3D talking heads.In order to accelerate the training process, we opted to train the two models independently. This approach resulted in improved convergence for both models.

$$L_{S2L} = \lambda_1 L_{rec} + \lambda_2 L_{mouth} + \boxed{\lambda_3 L_{cos}} + \lambda_4 L_{vel}.$$

$$L_{S2D} = \lambda_1 L_{rec} + \boxed{\lambda_2 L_{cos}} + \lambda_3 L_{weighted}.$$

Incorporating the **Cosine Loss** (L_{Cos}) during training of the two models enhances the fidelity of the generated displacements for both landmarks and vertices. We evaluate of our proposed framework in comparison to **Faceformer** [2] and **VOCA** [1].

Results


Trespassing is forbidden and subject to penalty.
A photograph can become a piece of art.
Sentence
She had your dark suit in greasy wash water all year.
Time
0
0.5
1
1.5
2
2.5
3
GT
VOCA
FaceFormer
Ours

	Landmarks			Dense		
Methods	**LE** (mm)	**DE** (mm)	**DAE** (Rad)	**LE** (mm)	**DE** (mm)	**DAE** (Rad)
VOCA [1]	0.87	0.77	0.29	0.72	0.62	0.23
Faceformer [2]	0.61	0.56	0.20	0.51	0.42	0.17
Ours	**0.50**	**0.44**	**0.13**	**0.43**	**0.34**	**0.12**

	Landmarks			Dense		
Loss	**LE** (mm)	**DE** (mm)	**DAE** (Rad)	**LE** (mm)	**DE** (mm)	**DAE** (Rad)
w/o L_{cos}	0.53	0.46	0.19	0.44	0.36	0.17
w/ L_{cos}	**0.50**	**0.44**	**0.13**	**0.43**	**0.34**	**0.12**

References

[1] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, Michael J. Black: Capture, Learning, and Synthesis of 3D Speaking Styles. CoRR abs/1905.03079 (2019)

[2] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, Taku Komura: FaceFormer: Speech-Driven 3D Facial Animation with Transformers. CoRR abs/2112.05329 (2021)

Look at some results:

Vote here for poster: