MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement

Gege Shan

Abstract

This paper presents a generic method for generating full facial 3D animation from speech. Existing approaches to audio-driven facial animation exhibit uncanny or static upper face animation, fail to produce accurate and plausible co-articulation or rely on person-specific models that limit their scalability. To improve upon existing models, MeshTalk propose a generic audio-driven facial animation approach that achieves highly realistic motion synthesis results for the entire face. At the core of the approach is a categorical latent space for facial animation that disentangles audiocorrelated and audio-uncorrelated information based on a novel cross-modality loss. The approach ensures highly accurate lip motion, while also synthesizing plausible animation of the parts of the face that are uncorrelated to the audio signal, such as eye blinks and eye brow motion. The approach outperforms several baselines and obtains state-of-the-art quality both qualitatively and quantitatively.

Keywords：3D Face Animation, Latent Space, Audio.

1 Introduction

Speech-driven facial animation is a highly challenging research problem with several applications such as facial animation for computer games, e-commerce, or immersive VR telepresence. The demands on speech-driven facial animation differ depending on the application. Applications such as speech therapy or entertainment (e.g., Animojies or AR effects) do not require a very presice level of realism in the animation. In the production of films, movie dubbing, driven virtual avatars for e-commerce applications or immersive telepresence, on the contrary, the quality of speech animation requires a high degree of naturalness, plausibility, and has to provide intelligibility comparable to a natural speaker. The human visual system has been evolutionary adapted to understanding subtle facial motions and expressions. Thus, a poorly animated face without realistic coarticulation effects or out of lip-sync is deemed to be disturbing for the user.

Psychological literature has observed that there is an important degree of dependency between speech and facial gestures. This dependency has been exploited by audiodriven facial animation methods developed in computer vision and graphics[1]. With the advances in deep learning, recent audio-driven face animation techniques make use of person-specific approaches[2] that are trained in a supervised fashion based on a large corpus of paired audio and mesh data. These approaches are able to obtain high-quality lip animation and synthesize plausible upper face motion from audio alone. To obtain the required training data, high-quality vision-based motion capture of the user is required, which renders these approaches as highly impractical for consumer-facing applications in real world settings. Recently, Cudeiro et al.[3] extended this work, by proposing

a method that is able to generalize across different identities and is thus able to animate arbitrary users based on a given audio stream and a static neutral 3D scan of the user. While such approaches are more practical in real world settings, they normally exhibit uncanny or static upper face animation[3].

The reason for this is that audio does not encode all aspects of the facial expressions, thus the audiodriven facial animation problem tries to learn a one-to-many mapping, i.e., there are multiple plausible outputs for every input. This often leads to over-smoothed results, especially in the regions of the face that are only weakly or even uncorrelated to the audio signal.

This paper proposes a novel audio-driven facial animation approach that enables highly realistic motion synthesis for the entire face and also generalizes to unseen identities. This paper learns a novel categorical latent space of facial animation that disentangles audio-correlated and audio-uncorrelated information, e.g., eye closure should not be bound to a specific lip shape. The latent space is trained based on a novel cross-modality loss that encourages the model to have an accurate upper face reconstruction independent of the audio input and accurate mouth area that only depends on the provided audio input. This disentangles the motion of the lower and upper face region and prevents over-smoothed results. Motion synthesis is based on an autoregressive sampling strategy of the audio-conditioned temporal model over the learnt categorical latent space. This approach ensures highly accurate lip motion, while also being able to sample plausible animations of parts of the face that are uncorrelated to the audio signal, such as eye blinks and eye brow motion. The contributions of the reproduced papers are as follows:

• A novel categorical latent space for facial animation synthesis that enables highly realistic animation of the whole face by disentanglement of the upper and lower face region based on a cross-modality loss.

• An autoregressive sampling strategy for motion synthesis from an audio-conditioned temporal model over the learned categorical latent space.

• This approach outperforms the current state-of-the-art both qualitatively and quantitatively in terms of obtained realism.

2 Related works

Speech-based face animation has a long history in computer vision and ranges from artist-friendly stylized and viseme-based models[4-5] to neural synthesis of 2D[2] and 3D[2,6-7] faces. In the following, we review the most relevant approaches.

Viseme-based face animation. In early approaches, a viseme sequence is generated from input text or directly from speech using HMM-based acoustic models. Visual synthesis is achieved by blending between face images from a database. For 3D face animation, Kalberer et al. model viseme deformations via independent component analysis. In[8], 3D face masks are projected onto a low-dimensional eigenspace and smooth temporal animation is achieved by spline fitting on each component of this space, given a sequence of key viseme masks. Leaning on the concept on phonetic co-articulation, Martino et al. seek to find context-dependent visemes rather than blending between key templates. Given the success of JALI[4], an animator-centric audio-

drivable jaw and lip model, Zhou et al.[5] propose an LSTM-based, near real-time approach to drive an appealing lower face lip model. Due to their generic nature and artist-friendly design, viseme based approaches are popular for commercial applications particularly in virtual reality.

Speech-driven 2D talking heads. Many speech-driven approaches are aimed at generating realistic 2D video of talking heads. Early work replaced the problem of learning by searching in existing video for similar utterances as the new speech. Brand et al. proposed a generic ML model to drive a facial control model that incorporates vocal and facial dynamic effects such as co-articulation. The approach of Suwajanakorn et al. is able to generate video of a single person with accurate lip sync by synthesizing matching mouth textures and compositing them on top of a target video clip. However, this approach only synthesizes the mouth region and requires a large corpus of personalized training data (≈17 hours). Wav2lip tackles the problem of visual dubbing, i.e., of lip-syncing a talking head video of an arbitrary person to match a target speech segment. Neural Voice Puppetry performs audio-driven facial video synthesis via neural rendering to generate photo-realistic output frames. X2Face is an encoder/decoder approach for 2D face animation, e.g., from audio, that can be trained fully self-supervised using a large collection of videos. Other talking face video techniques are based on generative adversarial networks (GANs). The approach of Vougioukas et al. synthesizes new upper face motion, but the results are of low resolution and look uncanny. The lower face animation approach of Chung et al. requires only a few still images of the target actor and a speech snippet as input. To achieve this, an encoder-decoder model is employed that discovers a joint embedding of the face and audio. Zhou et al. improve on Chung et al. by learning an audio-visual latent space in combination with an adversarial loss that allows to synthesize 2D talking heads from either video or audio. All the described 2D approaches operate in pixel space and can not be easily generalized to 3D.

Speech-driven 3D models. Approaches to drive 3D face models mostly use visual input. While earlier works map from motion captures or 2D video to 3D blendshape models[9-11], more recent works provide solutions to animate photo-realistic 3D avatars using sensors on a VR headset[7,12]. These approaches achieve highly realistic results, but they are typically personalized and are not audio-driven. Most fully speech-driven 3D face animation techniques require either personalized models[2,7,13] or map to lower fidelity blendshape models or facial landmarks. Cao et al.[13] propose speech-driven animation of a realistic textured personalized 3D face model that requires mocap data from the person to be animated, offline processing and blending of motion snippets. The fully speech-driven approach of Richard et al. enables realtime photo-realistic avatars, but is personalized and relies on hours of training data from a single subject. Karras et al. learn a speech-driven 3D face mesh from as little as 3-5 minutes of data per subject and condition their model on emotion states that lead to facial expressions. In contrast to our approach, however, this model has lower fidelity lip sync and upper face expressions, and does not generalize to new subjects.In[6], a single-speaker model is generalized via re-targeting techniques to arbitrary stylized avatars. Most closely related to our approach is VOCA[3], which allows to animate arbitrary neutral face meshes from audio and achieves convincing lip synchronization. While generating appealing lip motion, their model can not synthesize upper face motion and tends to generate muted

expressions. Moreover, the approach expects a training identity as conditional input to the model. As shown by the authors in their supplemental video, this identity code has a high impact on the quality of the generated lip synchronization. Consequentially, MeshTalk found VOCA to struggle on large scale datasets with hundreds of training subjects. In contrast to the discussed works, our approach is non-personalized, generates realistic upper face motion, and leads to highly accurate lip synchronization.

3 Method

Our goal is to animate an arbitrary neutral face mesh using only speech. Since speech does not encode all aspects of the facial expressions –eye-blinks are a simple example of uncorrelated expressive information – most existing audiodriven approaches exhibit uncanny or static upper face animation[3]. To overcome this issue, MeshTalk learn a categorical latent space for facial expressions. At inference time, an autoregressive sampling from a speech-conditioned temporal model over this latent space ensures accurate lip motion while synthesizing plausible animation of face parts that are uncorrelated to speech. With this in mind, the latent space should have the following properties:Categorical. Most successful temporal models operate on categorical spaces[14-15]. In order to use such models, the latent expression space should be categorical as well.

Expressive. The latent space must be capable of encoding diverse facial expressions, including sparse events like eye blinks.

Semantically disentangled. Speech-correlated and speechuncorrelated information should be at least partially disentangled, e.g., eye closure should not be bound to a specific lip shape.

3.1 Overview

The network architecture of the system is shown in Figure1. A series of animated face grids (expression signals) and speech signals are mapped to a categorized latent expression space. A UNet-style decoder is then used to animate a given grid of neutral face templates based on the encoded expressions. The audio encoder is a four-layer, one-dimensional temporal convolutional network, similar to the one used in[7]. The expression encoder has three fully connected layers followed by an LSTM layer to capture temporal dependencies. The fusion module is a three-layer MLP. decoder D has a UNet-style architecture with additive skip connections. this structural inductive bias prevents the network from diverging too much from the given template mesh. At the bottleneck layer, the expression code c1:T,1:H is connected to the encoded template mesh. The bottleneck layer is followed by two LSTM layers to model the temporal dependence between frames, and then three fully connected layers to remap the representation to the vertex space.

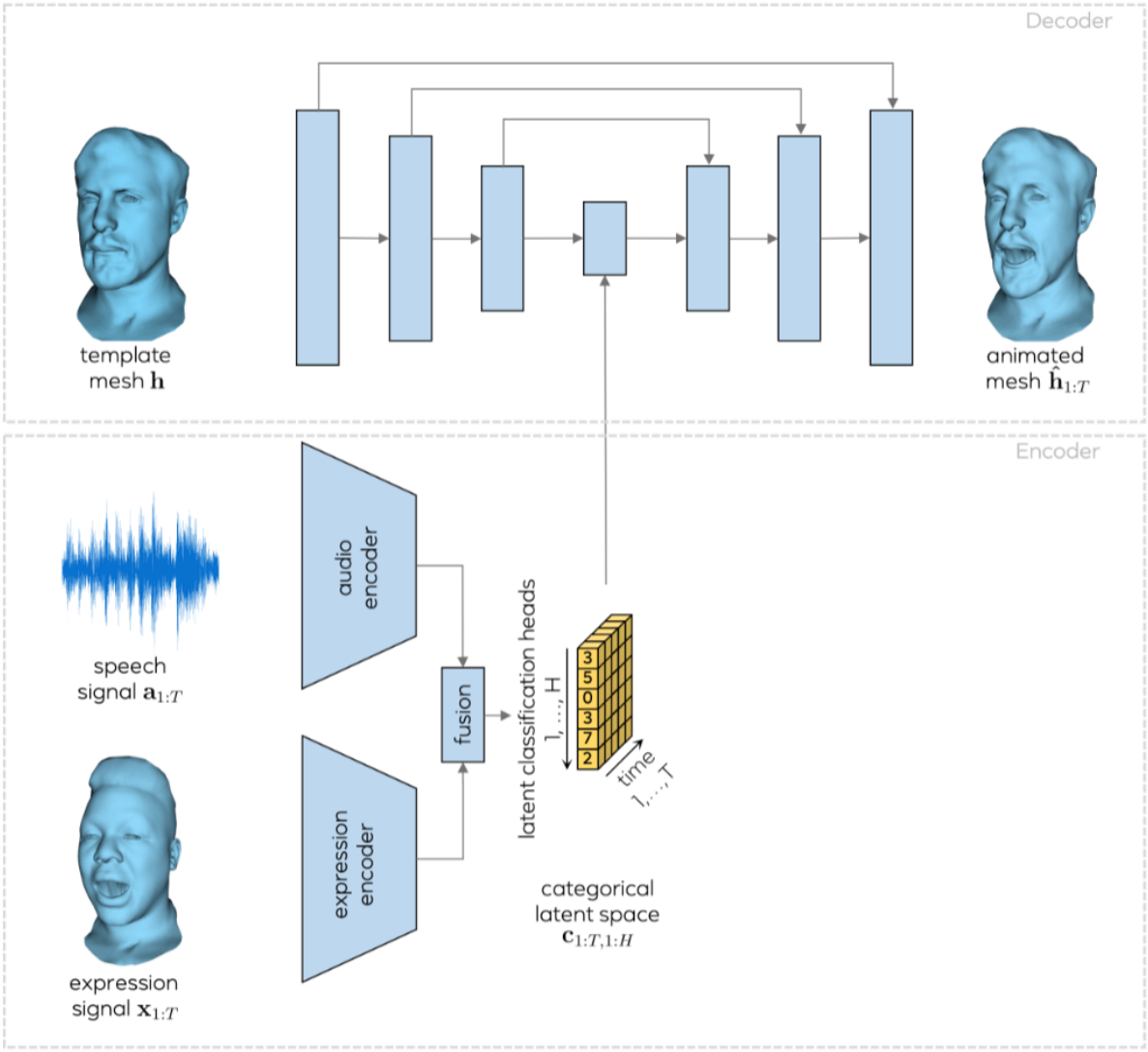

Figure 1: Overview of the system

3.2 Loss

This paper proposes a novel loss function:

$$\mathcal{L}_{xMod} = \sum_{t=1}^{T}\sum_{v=1}^{V} \mathcal{M}_v^{(upper)}\left(\left\|\hat{h}_{t,v}^{(expr)} - x_{t,v}\right\|^2\right) + \sum_{t=1}^{T}\sum_{v=1}^{V} \mathcal{M}_v^{(mouth)}\left(\left\|\hat{h}_{t,v}^{(audio)} - x_{t,v}\right\|^2\right) \tag{1}$$

$$\mathcal{L}_{eyelid} = \sum_{t=1}^{T}\sum_{v=1}^{V} \mathcal{M}_v^{(eyelid)}\left(\left\|\hat{h}_{t,v} - x_{t,v}\right\|^2\right) \tag{2}$$

where $\mathcal{M}^{(upper)}$ is a mask that assigns a high weight to vertices on the upper face and a low weight to vertices around the mouth. Similarly, $\mathcal{M}^{(mouth)}$ assigns a high weight to vertices around the mouth and a low weight to other vertices. The cross-modality loss encourages the model to have an accurate upper face reconstruction independent of the audio input and, accordingly, to have an accurate reconstruction of the mouth area based on audio independent of the expression sequence that is provided. Since eye blinks are quick and sparse events that affect only a few vertices, MeshTalk also found it crucial to emphasize the loss on the eye lid vertices during training. MeshTalk therefore adds a specific eye lid loss. where $\mathcal{M}^{(eyelid)}$ is a binary mask with ones for eye lid vertices and zeros for all other vertices. The final loss this paper optimize is $\mathcal{L} = \mathcal{L}_{xMod} + \mathcal{L}_{eyelid}$. This paper found that an equal weighting of the two terms works well in practice.

4 Implementation details

4.1 Comparing with released source codes

A big problem of speech-based generation of 3D facial faces is that the generated facial expressions are not rich enough, and the problem may be effectively alleviated by a multimodal approach. The audio is input during the training of the model, and the model uses automatic speech recognition to obtain the speaking text, together with the emotional labels of the audio and text and the 3D model of the speaker, to finally generate a 3D talking head with vivid expressions. This is the improvement idea of this paper, and the improvement has not been completed yet due to the time problem.

5 Conclusion and future work

This paper has presented a generic method for generating 3D facial animation from audio input alone. A novel categorical latent space in combination with a cross-modality loss enables autoregressive generation of highly realistic animation. This approach has demonstrated highly accurate lip motion, while also synthesizing plausible motion of uncorrelated regions of the face. It outperforms several baselines and obtains state-of-the-art quality. This approach can be a stepping stone towards VR telepresence applications. Future work will likely focus on detailed reconstruction of 3D talking head.

References

[1] BREGLER C, COVELL M, SLANEY M. Video rewrite: Driving visual speech with audio[C]// Proceedings of the 24th annual conference on Computer graphics and interactive techniques. 1997: 353-360.

[2] KARRAS T, AILA T, LAINE S, et al. Audio-driven facial animation by joint end-to-end learning of pose and emotion[J]. ACM Transactions on Graphics (TOG), 2017, 36(4): 1-12.

[3] CUDEIRO D, BOLKART T, LAIDLAW C, et al. Capture, learning, and synthesis of 3D speaking styles [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 10101-10111.

[4] EDWARDS P, LANDRETH C, FIUME E, et al. Jali: an animator-centric viseme model for expressive lip synchronization[J]. ACM Transactions on graphics (TOG), 2016, 35(4): 1-11.

[5] ZHOU Y, XU Z, LANDRETH C, et al. Visemenet: Audio-driven animator-centric speech animation[J]. ACM Transactions on Graphics (TOG), 2018, 37(4): 1-10.

[6] TAYLOR S, KIM T, YUE Y, et al. A deep learning approach for generalized speech animation[J]. ACM Transactions On Graphics (TOG), 2017, 36(4): 1-11.

[7] RICHARD A, LEA C, MA S, et al. Audio-and gaze-driven facial animation of codec avatars[C]// Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021: 41-50.

[8] KALBERER G A, VAN GOOL L. Face animation based on observed 3d speech dynamics[C]//Proceedings Computer Animation 2001. Fourteenth Conference on Computer Animation (Cat. No. 01TH8596). 2001: 20-251.

[9] DENG Z, CHIANG P Y, FOX P, et al. Animating blendshape faces by cross-mapping motion capture data[C]//Proceedings of the 2006 symposium on Interactive 3D graphics and games. 2006: 43-48.

[10] WANG L, HAN W, SOONG F K. High quality lip-sync animation for 3D photo-realistic talking head[C]//2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2012: 4529-4532.

[11] GARRIDO P, ZOLLHÖFER M, CASAS D, et al. Reconstruction of personalized 3D face rigs from monocular video[J]. ACM Transactions on Graphics (TOG), 2016, 35(3): 1-15.

[12] WEI S E, SARAGIH J, SIMON T, et al. Vr facial animation via multiview image translation[J]. ACM Transactions on Graphics (TOG), 2019, 38(4): 1-16.

[13] CAO Y, TIEN W C, FALOUTSOS P, et al. Expressive speech-driven facial animation[J]. ACM Transactions on Graphics (TOG), 2005, 24(4): 1283-1302.

[14] VAN DEN OORD A, KALCHBRENNER N, ESPEHOLT L, et al. Conditional image generation with pixelcnn decoders[J]. Advances in neural information processing systems, 2016, 29.

[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.