Speech-based 3D Face Animation

Abdullah Ahmed Ankur Aditya Alejandro Rodriguez Pascual
University of Massachusetts Amherst
Amherst, MA 01003
{ amahmed, aaditya, arodriguezpa }@umass.edu

Abstract

This paper implements a simpler version of MeshTalk for generating 3D animated face from speech. Existing approaches mainly focuses on upper face animation leading to it's limitation in their scalability. With the MeshTalk we aim to not only animate the lips movement for a given speech but also focus on upper face part like eyebrows which are uncorrelated with the speech signal. The fundamental foundation of MeshTalk methodology revolves around the utilization of a categorical latent space specifically tailored for facial animation purposes. This innovative latent space possesses the remarkable capability to disentangle information into two distinct categories: audio-correlated and audio-uncorrelated. The distinguishing factor lies in the novel cross-modality loss, enabling enhanced control over facial animation outcomes. For training, we use the dataset from an open source 'multiface' dataset containing data of two unique users/ mesh .

1. Introduction

Speech-based 3D face animation is a highly challenging problem, but, at the same time, very demanding due to its applications in video games, virtual reality, and augmented reality. This advancement in 3D face animation technology opens up new possibilities for creating lifelike characters and immersive virtual experiences. In applications like movie dubbing, speech animation must be natural, believable and easy to understand. The human visual system is intelligent enough to understand the subtle facial emotions and expression. Hence there is a huge demand in research which aims at at filling the gap capturing the whole facial expressions instead of only the upper face or lower face.

Capturing the many intricacies in mouth movements during speech can prove too complicated or tedious for most, if not all, animators. It becomes especially difficult when similar animations must be applied to different faces. If a high-quality animation can be applied to one face, it stands to reason that the same animation can be mapped from that face to other faces, replicating similar movements in a way that still fits the new geometry. A machine learning approach seems appropriate for this problem, as we can model mouth movements relative to a neutral face expression, and ideally transfer those movements to other faces with similar results. Speech can help guide these mappings through additional information given by the audio created by what the person is saying with a given face expression.

This paper focuses on speech based 3D face animation approach which aims to generate naturalistic facial expressions. To achieve this we make use of MeshTalk's novel categoircal latent space for learning that disentangles audio-correlated and audio-uncorrelated information, e.g., eyebrow movement should not depend on a particular lip movement. This latent space is trained based on the cross-modality loss which aims to have an accurate upper face animation irrespective of the speech input and accurate lower face expressions which is dependent on the speech input.

For training we leverage the multiface dataset by Facebook Research.

2. Related Work

In this section we review the most relevant works in the field of speech based 3D face animation.

Facebook Research's MeshTalk achieves this same purpose with surprising performance [3]. Through a dual encoder that processes one face's position together with an audio snippet, combines the resulting mapping, and applies it to another face's encoded neutral position to apply the same face position to the new face after decoding.

Learning Audio-Driven Viseme Dynamics for 3D Face Animation introduces an algorithm that maps phonemes to viseme curves in a way that can be transferred between languages, allowing audio-to-face mappings at their most fundamental level. These viseme curves can then be applied to multiple different face models to achieve similar animations throughout different faces [1].

FaceFormer: Speech-Driven 3D Facial Animation with Transformers provides a unique autoregressive approach using transformers to iteratively map a piece of audio to a

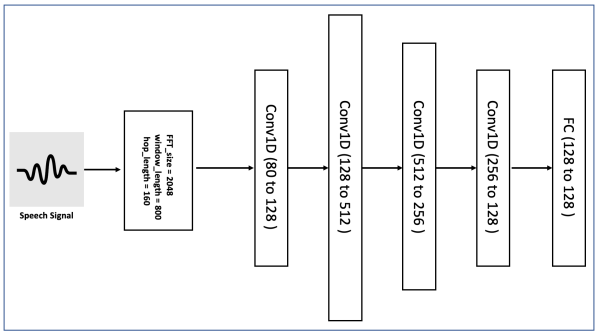

Figure 1. Audio Encoder

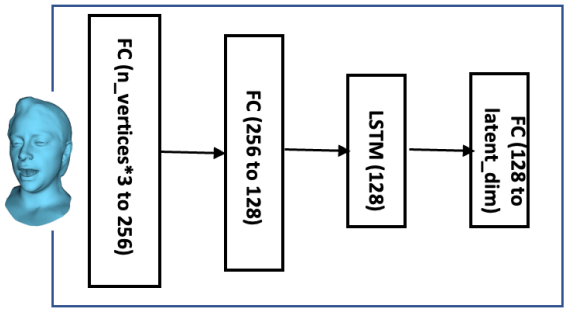

Figure 2. Expression Encoder

face's expression, starting from a face with a neutral expression [2].

3. Method

We simplified MeshTalk and made a few modifications to compare performance. We implemented a dual encoder architecture that encodes an audio sample on one hand, and the corresponding facial expression 3D mesh on the other. Then, the encoded output from both encoders is concatenated and fed to a MLP to combine the two encoded outputs. Finally, the joint encoding is applied to the target face through an encoder-decoder network.

3.1. Audio Encoder

The audio encoder (Fig. 1) converts the audio file into an 80-dimensional vector. This vector is fed to a convolutional layer to expand it to 128 dimensions. After that, the resulting vector is fed through 3 convolutional layers to encode it into a 128-dimensional latent space, using Xavier uniform initialization for every layer. Finally, the resulting vector is fed to a fully-connected layer that delivers the final 128-dimensional audio encoding. We also implemented a leaky ReLU layer after every convolutional layer, and a dropout layer after each of the last 3 convolutional layers.

3.2. Expression Encoder

The expression encoder (Fig. 2) consists of 3 linear layers and a one-layer RNN with LSTM. This encodes the mesh of a face with an expression matching the audio sample into a latent 128-dimensional space. We implemented a leaky ReLU layer after the first and second linear layers.

3.3. Decoder

The joint encoder consists of three linear layers that encode the concatenated output of the audio and expression encoders into a vector assigning probabilities of the which preset head it is and which expression it is representing. We also implemented leaky ReLU layers in between the linear layers.

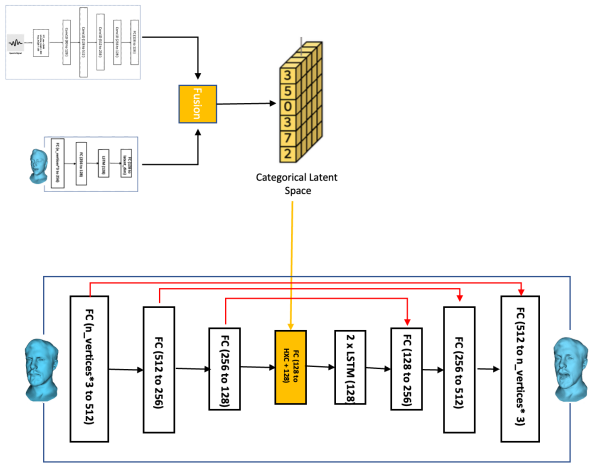

Figure 3. Decoder Architecture

The encoder-decoder (Fig. 3) network consists of a three layer encoder that encodes target face into a 128-dimensional latent embedding, and a 3-layer decoder that decodes the combined encodings into a mesh for the target face with the corresponding expression. The encoder consists of 3 linear layers with leaky ReLU layers applied afterwards. Then, the resulting encoding is concatenated then fused with the joint encoding with a linear layer and a leaky ReLU layer applied afterwards. After that, the fused encoding is fed to a one-layer RNN with LSTM. Finally, the resulting 128-dimensional vector is fed to the decoder, which consists of 3 linear layers with leaky ReLU layers applied after them. Additionally, target face encodings of the same dimensionality are added to the decodings as they pass through the decoder (i.e.: the encoding after one layer of the encoder is applied to the decoding after 3 layers of the decoder, since they have the same dimensions), in order to enforce that the encoded data is fit to the target face's geometry.

3.4. Autoregressive model

Audio conditioned autoregressive model (Fig. 4) is used to learn the learn the temporal model over the categorical latent space. It's used as while driving the template mesh with audio only input, the expression input is not available. Hence the missing information that cannot be pulled from the audio has to be synthesised. This model thereby allows us to generate possible expression so as it is consistent with the audio input.

Autoregressive model is a temporal convolutional network with four layers. Each of these layers have the kernel size of 2 and temporal dilations of 1, 2, 4 and 8 and 64 channels per categorical head. Convolutions in each of these layers are masked such that only past temporal information from previous categorical heads is visible in each layer. Also each of these layer is conditioned on the audio which is then concatenated with the remaining layer input. It is worth noting that the audio encoder (Fig. 1) remains fixed when we train the autoregressive model.

3.5. Training Losses

Let the $\mathbf{x}_{1:T} = (\mathbf{x}_1, \ldots, \mathbf{x}_T), \mathbf{x}_t \in \mathcal{R}^{V\times 3}$ be a sequence of T face meshes, each of them represented by V vertices. Let $\mathbf{a}_{1:T} = (\mathbf{a}_1, \ldots, \mathbf{a}_T), \mathbf{a}_t \in \mathcal{R}^D$ be a sequence of T speech sinppets wach with D samples. The template mesh is represented as $\mathbf{h} \in \mathcal{R}^{V\times 3}$.

One of aim is to achieve high expressiveness of the categorical latent space, for which the space must be sufficiently large. Hence to achieve that the we model H latent classification heads of C-way categorical, allowing the large expression space for small number of categories.Throughout our implementation we have set C = 128 and Heads =164.

Encoder ϵ maps the expression and audio sequences to multi-head categorical latent space $(T \times H \times C)$ dimensional encoding

$$\mathbf{e}_{1:T,1:H,1:C} = \epsilon(\mathbf{x}_{1:T}, \mathbf{a}_{1:T}) \in \mathcal{R}^{T\times H\times C} \tag{1}$$

This encoding is then transformed to a Gumbel-Softmax over each latent classification head (2). Then the animation of output template mesh $\mathbf{h}$ is realized by the decoder $\mathcal{D}$ (3),

$$\mathbf{c}_{1:T,1:H} = [Gumbel(\mathbf{e}_{1:T,1:H,1:C})]_{1:T,1:H} \tag{2}$$

$$\hat{\mathbf{h}}_{\mathbf{1:T}} = \mathcal{D}(\mathbf{h}, \mathbf{c}_{1:T,1:H}) \tag{3}$$

For designing the cross-modality loss function, we need to ensure that the information from both the input modalities is utilized in the latent space. If we simply train the l_2 reconstruction loss between input template mesh $\hat{\mathbf{h}}_{\mathbf{1:T}}$ and expression input, $\mathbf{x}_{1:T}$, it would lead to the audio input to be completely ignored as expression signal already contains enough information for reconstruction. Hence to tackle this let h_x represent the template mesh for the person represented in the expression signal $\mathbf{x}_{1:T}$. Instead of single reconstructions, two different reconstruction are generated.

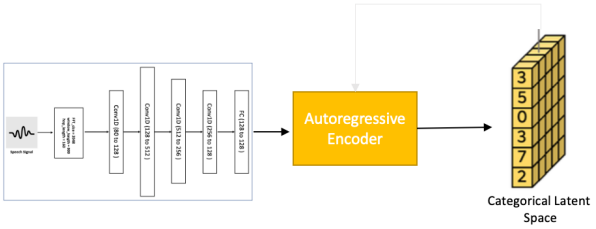

Figure 4. Autoregressive Model

$$\hat{\mathbf{h}}^{\mathbf{audio}}_{\mathbf{1:T}} = \mathcal{D}(\mathbf{h}_x, \epsilon(\tilde{\mathbf{x}}_{1:T}, \mathbf{a}_{1:T})) \tag{4}$$

$$\hat{\mathbf{h}}^{\mathbf{expr}}_{\mathbf{1:T}} = \mathcal{D}(\mathbf{h}_x, \epsilon(\mathbf{x}_{1:T}, \tilde{\mathbf{a}}_{1:T})) \tag{5}$$

where, $\tilde{\mathbf{x}}_{1:T}$ and $\tilde{\mathbf{a}}_{1:T}$) are a randomly sampled expression and audio sequences from training set. Hence the cross-modality function is defined as:

$$\begin{aligned}\mathcal{L} = &\sum_{t=1}^{T}\sum_{v=1}^{V}\mathcal{M}_v^{(upper)}(||\hat{\mathbf{h}}_{t,v}^{expr} - \mathbf{x}_{t,v}||^2) \\ &+ \sum_{t=1}^{T}\sum_{v=1}^{V}\mathcal{M}_v^{(mouth)}(||\hat{\mathbf{h}}_{t,v}^{audio} - \mathbf{x}_{t,v}||^2) \\ &+ \sum_{t=1}^{T}\sum_{v=1}^{V}\mathcal{M}_v^{(eyelid)}(||\hat{\mathbf{h}}_{t,v} - \mathbf{x}_{t,v}||^2)\end{aligned} \tag{6}$$

where, $\mathcal{M}_v^{(upper)}$ and $\mathcal{M}_v^{(mouth)}$ are the masks that assigns higher weights to vertices on upper face and around the mouth respectively. , $\mathcal{M}_v^{(eyelid)}$ is a binary mask with 1 for eye-lid vertices, and 0 for rest.

4. Evaluation

4.1. Dataset

We train our model using a portion of facebook's multiface dataset [4]. Multiface contains 2D multi-view image data, tracked 3D mesh data, headpose data, and audio data of 13 participants, displaying different facial expressions and reading 50 phonetically balanced sentences, all captured through an array of 40 - 160 cameras at 30 frames per second. We use data 3D mesh data and corresponding audio data for training purposes. The mesh data is smoothed over by removing 1134 out of the 7,306 vertices. Audio, on the other hand, data is forced down to 16 KHz sampling rate.

4.2. Discussion & Results

Following the lead of the work in [3], we expected our results to mimic theirs. Accruing the benefits of both of their use of categorical, expressive latent space and the modality disentangling loss, Richard et al. were able to achieve higher lip syncing accuracy than other approaches while still better encoding (and exhibiting) upper facial expression and eyeblinks.

Given our model simplification and dataset summarization (described above in 4.1), it would have been natural observe an easier to train and run model at the cost overfitting specific template meshes and reduced mesh re-targeting functionality. We could not, however, complete our observations due to inability of running the training procedure on computer devices available to us. With 16 GB of memory on our personal work computers, running 3D mesh data was obvious impossibility. As such, we endeavored to perform training on computer GPUs. Despite the comparative surplus of memory available on GPUs, all attempts of training were eventually interrupted by memory allocation errors.

Furthermore, the 3D data handling library suggested to wrangle the mesh data by the authors (PyTorch3D) does not compile on any of the computing available to us, therefore preventing us from performing additional analysis on MeshTalk's results by examining and running their publicaly available codebase.

5. Conclusion

Leveraging categorical latent spaces, multimodal training, and feature disentanglement, the authors of MeshTalk avoid stale-expressioned face reconstructions while assuring high lip sunc accuarcy. Despite the great promise of the proposed technique, wrangling high amounts of dense 3D mesh data proved to be too difficult given constraints of a time limited course. This however motivates us to further study model and data distillation techniques for light weight, mobile networks for 3D mesh analysis and generation.

References

[1] Linchao Bao, Haoxian Zhang, Yue Qian, Tangli Xue, Changhai Chen, Xuefei Zhe, and Di Kang. Learning Audio-Driven Viseme Dynamics for 3D Face Animation. *arXiv*, 2023. 1

[2] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. FaceFormer: Speech-Driven 3D Facial Animation with Transformers. *arXiv*, 2021. 2

[3] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 00:1153–1162, 2021. 1, 4

[4] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. Multiface: A Dataset for Neural Face Rendering. *arXiv*, 2022. 3