Speech Based 3D Face Animation

Abdullah Ahmed Ankur Aditya

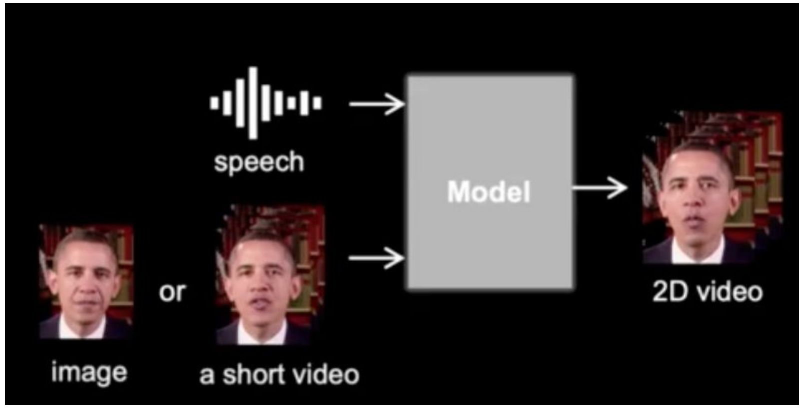
Alejandro Rodriguez Pascual

Introduction

- Speech-based 3D facial animation is a highly challenging problem
- In high demand
  - Video games
  - Virtual reality
  - Augmented reality
- Speech animation and lip-syncing is crucial to high-quality media
  - Otherwise, animation falls into the jarring “uncanny valley”
- As a result, there is high demand for cost-effective solutions
- An AI approach seems effective for this problem
  - Can map expressive faces to audio samples through encoding
  - Can use encodings to apply expressions to other faces
- Facebook Research’s MeshTalk achieves many of these goals, but is costly.

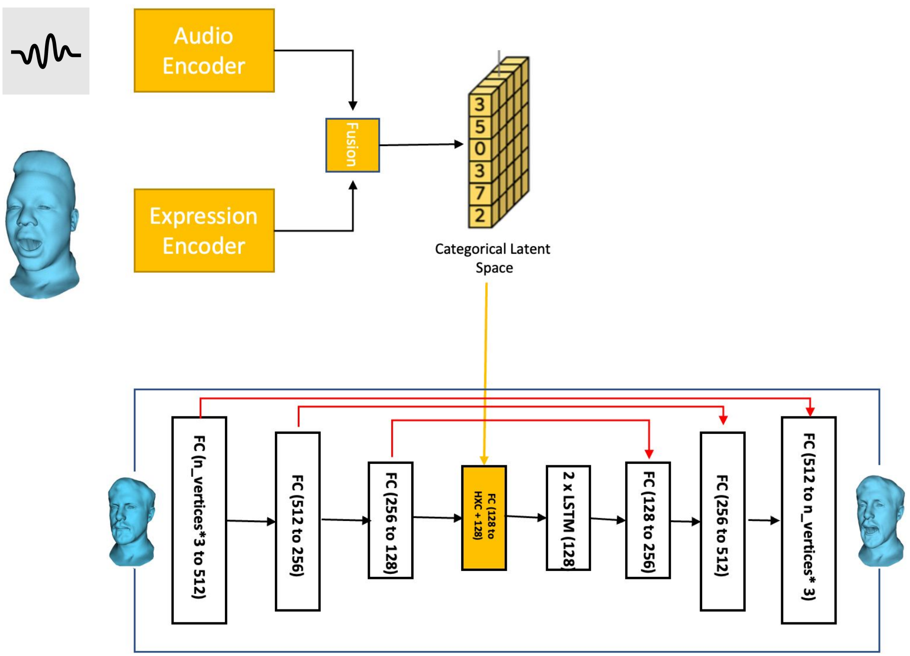
Background

2D Based Methods

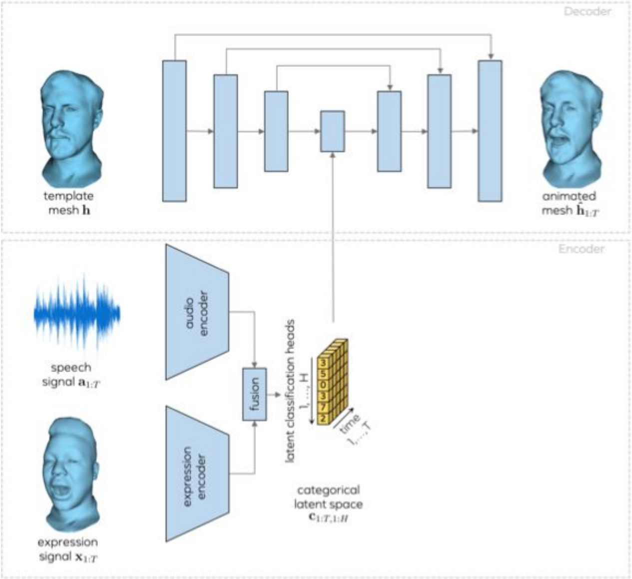


vs

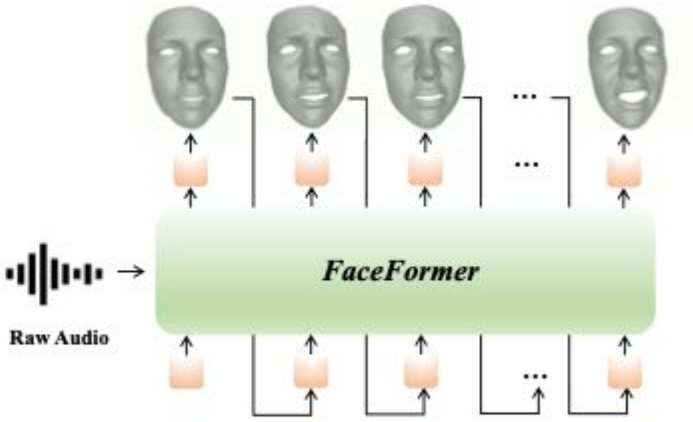
3D Mesh Based Methods



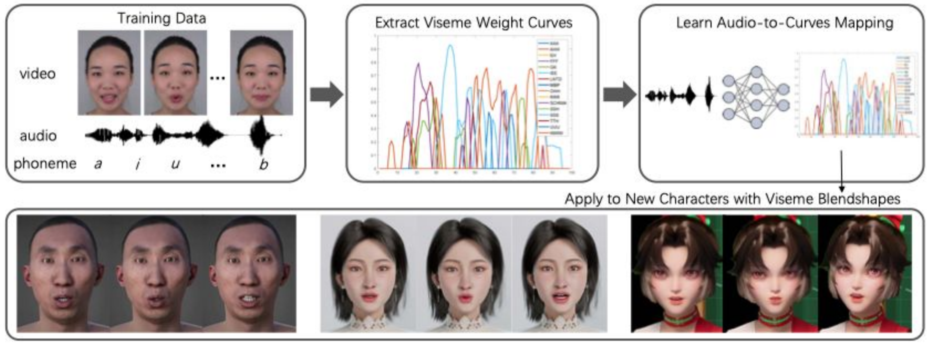
Related Works



Mesh Talk. [Richard et. al ICCV 2021]



FaceFormer [Yingruo et. al CVPR 2022]

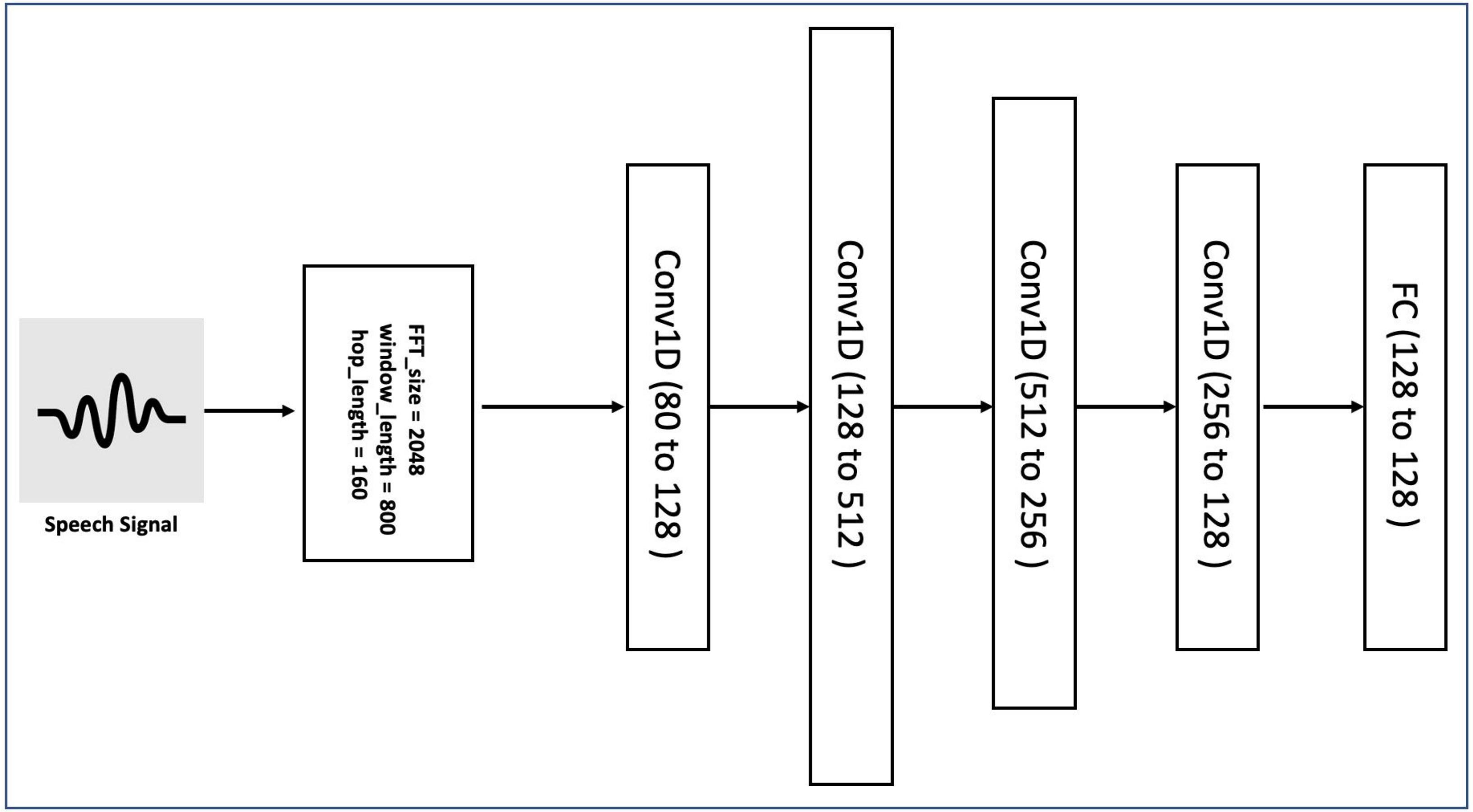


Audio Driven Viseme Dynamics
[Linchao et.]

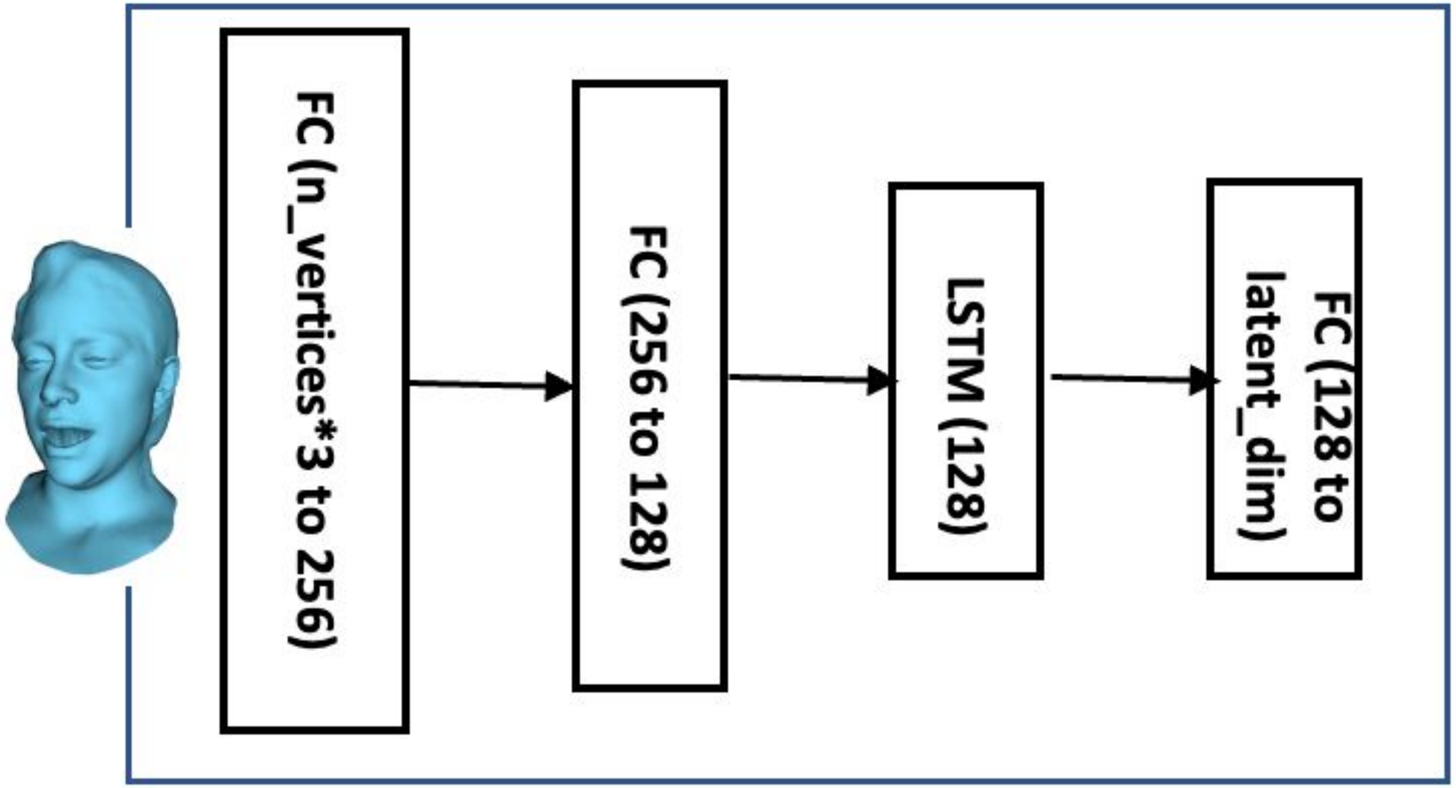
Approach

- We implemented a simplified version of Facebook Research's MeshTalk
  - Attempting to achieve similar results with reduced computational cost
  - Could make this solution more accessible and efficient
- Replaced some bloated architecture in favor of a simpler implementation
  - Reduced number of layers in some encoders
  - Streamlined loss calculations
- Made a few adjustments to try to retain accuracy
  - Increased layer size in some cases
  - Tracked models independently for optimization
- Simplified and reduced dataset to fit more limited hardware
  - Turned 20GB Multiface reduced dataset into <4GB dataset for training
  - Removed unnecessary data classes and reduced redundant data cases

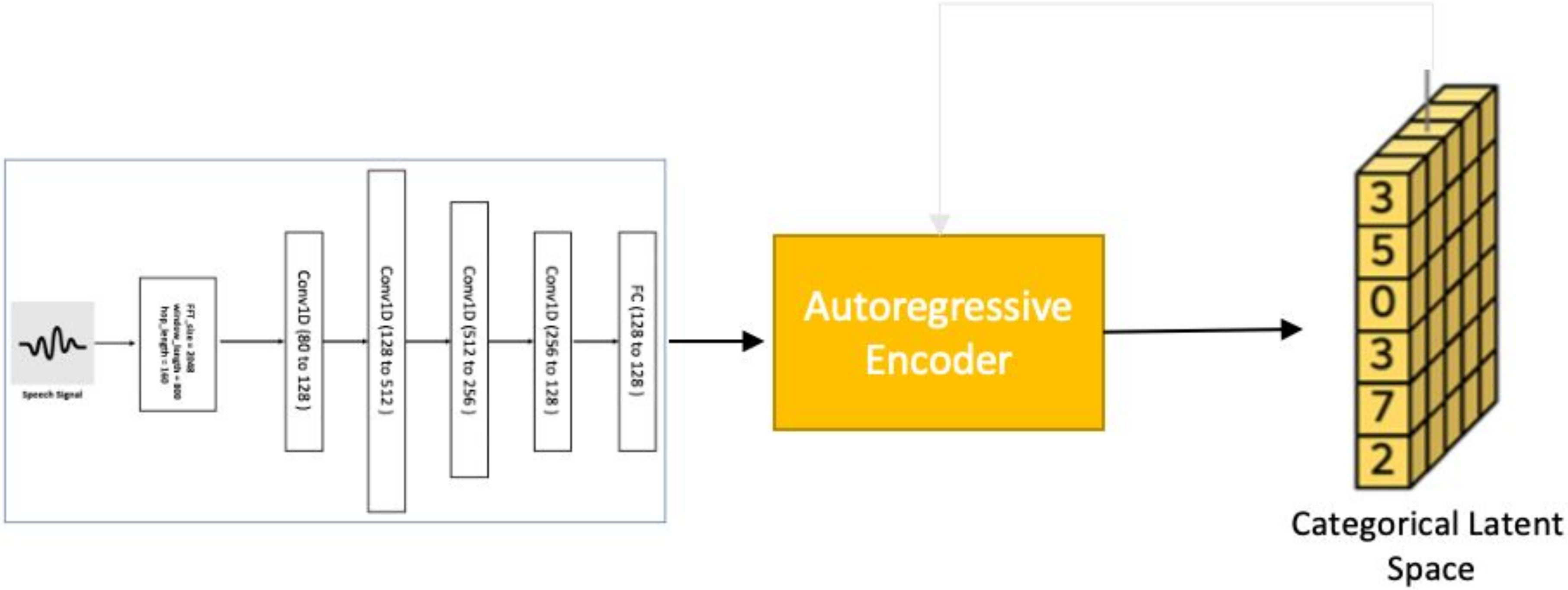
Audio Encoder



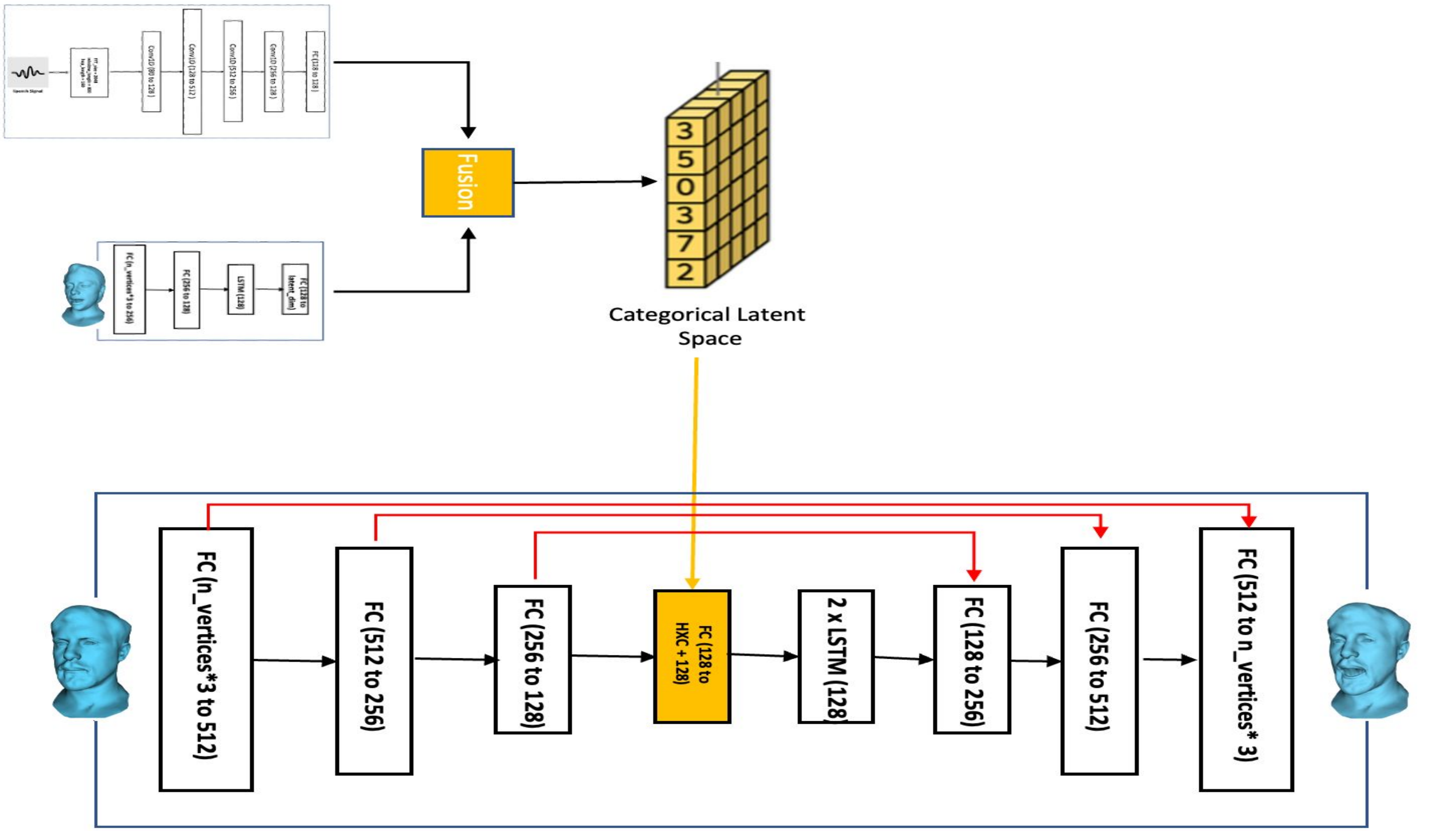
Expression Encoder



Auto-Regressive Model



Decoder Architecture



Training Loss

Step 1: Encoder ε maps the expression and audio sequences to multi-head categorical latent space

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

Step 2: This encoding is then transformed to a Gumbel-Softmax over each latent classification head. Then the animation of output template mesh h is realized by the decoder D.

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

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

$$\boxed{\begin{aligned}\hat{\mathbf{h}}^{\mathbf{audio}}_{\mathbf{1:T}} &= \mathcal{D}(\mathbf{h}_x, \epsilon(\tilde{\mathbf{x}}_{1:T}, \mathbf{a}_{1:T})) \\ \hat{\mathbf{h}}^{\mathbf{expr}}_{\mathbf{1:T}} &= \mathcal{D}(\mathbf{h}_x, \epsilon(\mathbf{x}_{1:T}, \tilde{\mathbf{a}}_{1:T}))\end{aligned}}$$

$\tilde{x}_{1:T}$ **: Randomly sampled expression sequence from training Set**

$\tilde{a}_{1:T}$ **: Randomly sampled audio sequence from training Set**

Cross-Modality loss function

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

(Expected) Results

- We were unable to train the model on available hardware
  - Error in a dump statement from an imported function when loading data
  - Nothing we can do to fix it, as we don’t own the code
- This challenge could have been overcome with better equipment/code
  - Memory limitations
  - Unoptimized code from an imported library
- Despite this setback, we are confident that our model could have achieved:
  - Significantly better runtime than MeshTalk due to the simplified approach
  - Noticeably worse accuracy due to our reduced dataset
  - Perhaps similar performance on a complete dataset
- It is hard to know for sure what the full results would have been without training and testing the model for ourselves

Conclusion

- MeshTalk shows a lot of promise in providing good 3D facial animation
  - Leverages categorical latent spaces, multimodal training, and feature disentanglement
  - Avoids stale-expressioned face reconstructions while assuring high lip sync accuracy.
- Very computationally costly, which we tried to correct
  - Simplified architecture
  - Reduced dataset
  - Streamlined code
- While we were unable to gather concrete results due to hardware limitations, we are confident that our approach could have provided key insights at worst, and a useful solution at best.