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

Anonymous ICCV submission

Paper ID 2644

	channels	kernel size	dilation
fully connected	128		
conv1D	128	5	2
conv1D	128	5	4
conv1D	128	5	6
conv1D	128	5	2
conv1D	128	5	4
conv1D	128	5	6
avg pool			
fully connected	128		

Table 1. Audio encoder architecture.

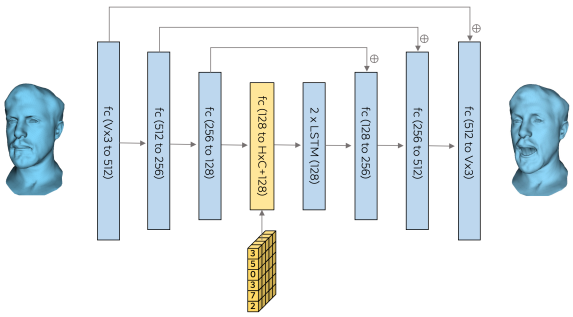

Figure 1. Decoder architecture.

1. Network Architecture Details

We provide details about the network architectures in this supplementary material. The activation functions used in all networks are leaky ReLUs.

1.1. Audio Encoder

The audio encoder first transforms the raw waveform into a Mel spectrogram with 80 frequency bins. We use an FFT size of $2,048$, window length of 800 samples, and a hop length of 160 samples, resulting in one new feature vector every 10ms. The network is a 1D temporal convolutional network with the architecture outlined in Table 1. We use skip connections between each layer and a dropout probability of 0.2.

1.2. Expression Encoder

The expression encoder flattens its $T \times V \times 3$ input mesh sequence to $T \times V \cdot 3$ and uses a fully connected layer to project the vertices from $V \cdot 3$ to 256 dimensions. This is followed by a second fully connected layer, mapping the representation further down to 128 dimensions. A single LSTM layer with 128 hidden units learns temporal information along the animated input mesh. The final expression code is a linear projection of the LSTM output to a 128 dimensional code.

1.3. Fusion Model

Audio and expression codes are concatenated and fused with three fully connected layers. The final output size of the fusion network is $T \times H \times C$, where T is the sequence length, H is the number of latent classification heads ($H = 64$), and C is the number of categories ($C = 128$). Categorization is achieved by a Gumbel softmax over the last dimension, such that the encoder output is a $T \times H$ tensor containing a categorical label in each component.

1.4. Decoder

The decoder (Figure 1) maps the template mesh to a 128-dimensional representation via three fully connected layers. This representation is concatenated with a one-hot representation of the categorical latent embedding and mapped back to 128 dimensions using a fully connected layer. Two LSTM layers with 128 hidden units each model temporal dependencies that arise from the latent embeddings. Finally, the representation is reprojected to the full $V \times 3$-dimensional vertex space. Note that all skip connections are purely additive to enforce that template mesh information can not be ignored.

1.5. Autoregressive Model

The autoregressive model is a temporal convolutional network with four layers, kernel size 2 in each layer, and temporal dilations of 1, 2, 4, and 8. Each layer has 64 channels per categorical head. Convolutions are masked such that only past temporal information and information from previous categorical heads is visible in each layer. Each layer is conditioned on an audio embedding that is concatenated to the remaining layer input. The audio embedding is obtained from the previously trained audio encoder (Table 1). We keep the audio encoder fix and do not fine-tune it when we train the autoregressive model.