READ Avatars: Realistic Emotion-controllable Audio Driven Avatars: Supplementary Material

Jack Saunders
https://jsaunders909.github.io/
Vinay Namboodiri
https://vinaypn.github.io/

Department of Computer Science
University of Bath
Bath, UK

In this supplementary material, we cover the more detailed model architecture, as well as the results of a user study for the ablation experiments.

1 Audio-To-Expression Network

The audio-to-expression network consists of a generator and discriminator.

The generator is given the 26 MFCC coefficients at three times the frame rate combined with the emotional label. The 7-dimensional emotional label (8 emotions with neutral as the zero vector) is repeated to match the MFCC coefficients, we use a window of 30 coefficients. This gives an input of shape $(30, 26+7=33)$. Fully connected layers map this from $(30,33) \rightarrow (30,64) \rightarrow (30,32)$. Each uses a LeakyReLU activation. Next, this is passed to a 1-layer, bidirectional LSTM to incorporate temporal information. This has shape $(30,64)$, we map this back to $(30,32)$ with a fully connected layer, then reduce the time dimension to match the frame rate of the video with max pooling along the temporal axis. Finally, fully connected layer map: $(10,32) \rightarrow (10,32) \rightarrow (10,64)$ again with LeakyRelu activation. The final layer then projects into the 103-dimensional parameter space (100 expressions + 2 eyelid + 1 jaw controls), this has no activation function.

Layer	Activation	Shape
Input	None	$(30,33)$
Linear	LeakyReLU	(30, 64)
Linear	LeakyReLU	(30, 32)
LSTM	None	(30, 64)
Linear	None	(30, 32)
Temporal Pool	None	(10, 32)
Linear	LeakyReLU	(30, 32)
Linear	LeakyReLU	(30, 64)
Linear	None	(30, 109)

Table 1: The layers of the audio-to-expression generator network. The discriminator follows a similar structure without temporal pooling.

The discriminator is very similar. The encoder now maps from the $(10,103)$-dimension parameter space that is the output of the generator to $(10,32)$ by LeakyReLU activated linear layers with dimensions 64 and 32. The LSTM is identical but this time we use no time pooling. The decoder uses linear layers of size 16 and 1.

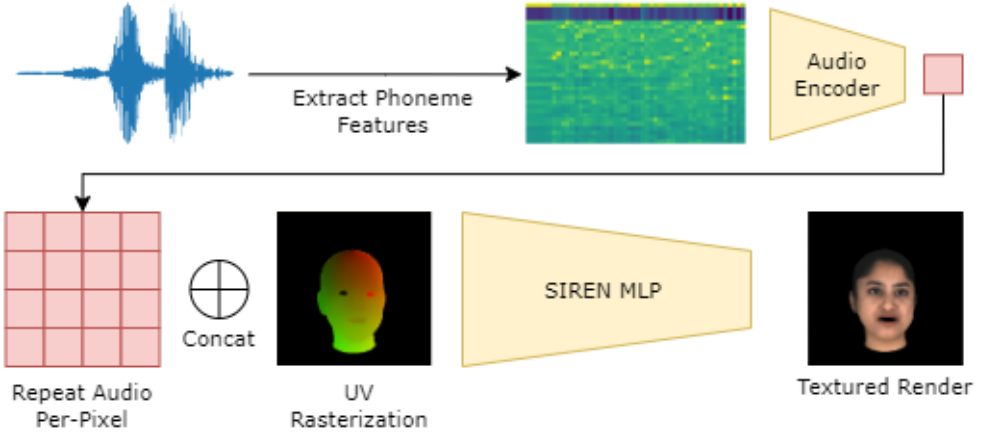

Figure 1: The structure of the audio-conditioned and resolution-independent SIREN neural texture.

2 Audio-Conditioned SIREN Neural Texture

Our neural texture model is based on the SIREN network. To condition it on audio, we first extract two per-frame phonemes as mentioned in the main paper. We take a window of 7 audio features centred on the current frame. This is encoded in a similar way to the audio in the audio-to-expression network, with fully connected layers mapping the now $(7,50)$ audio window to $(7,256)$ using 3 linear layers of size 256 with LeakyReLU activation. Instead of an LTSM we then use three temporal convolutions with kernel size 3 to get the same size output which is reshaped to $(1,7\times256)$ and is mapped to $(1,256)\rightarrow(1,7)$ using similar linear layers.

The ratserised uv image has shape $(256,256,2)$ and we repeat the encoded audio and concatenate it to the image to get $(256,256,9)$. Five 256-dimension sine-activated linear layers are applied to this, before a final, non-activated, linear layer converts this into a $(256,256,16)$ rasterized neural texture. This is shown in Figure 1.

3 Image-to-Image UNET

The image-to-image UNET takes a 16-channeled rasterized neural texture image, together with a window of 2-channel, rasterized uv images as input, and outputs the same window of predicted frames. To use all frames in the window, the frames are stacked in the channel dimension. We use 7 frames in each window, meaning that the input to the UNET is a $(7\times2)+16=30$ channel image, and the output is a $7\times3=21$ channel image. The UNET consists of 5 downsampling layers and 5 upsampling layers, connected with skip connections. The downsampling layers use stride 2 convolutions and the upsampling layers use bilinear interpolation.

Table 2: Results of the ablation study conducted with 30 users. Where the full method is preferred to the ablated version, we denote the result + and where the opposite is true, −

Ablation Statement	−	+
Ours > Ours w/o GAN	43.5	**56.5**
Ours > Ours w/o Audio Conditioning	44.0	**56.0**

4 Ablation User Study

Table 2 shows the result of an additional user study performed to further demonstrate the efficacy of our additional components. We asked 30 users, using Amazon's Mechanical Turk, if they preferred the results of our method with or without the named components. For both the GAN loss and the audio conditioning, the inclusion of our novel components was preferred overall.