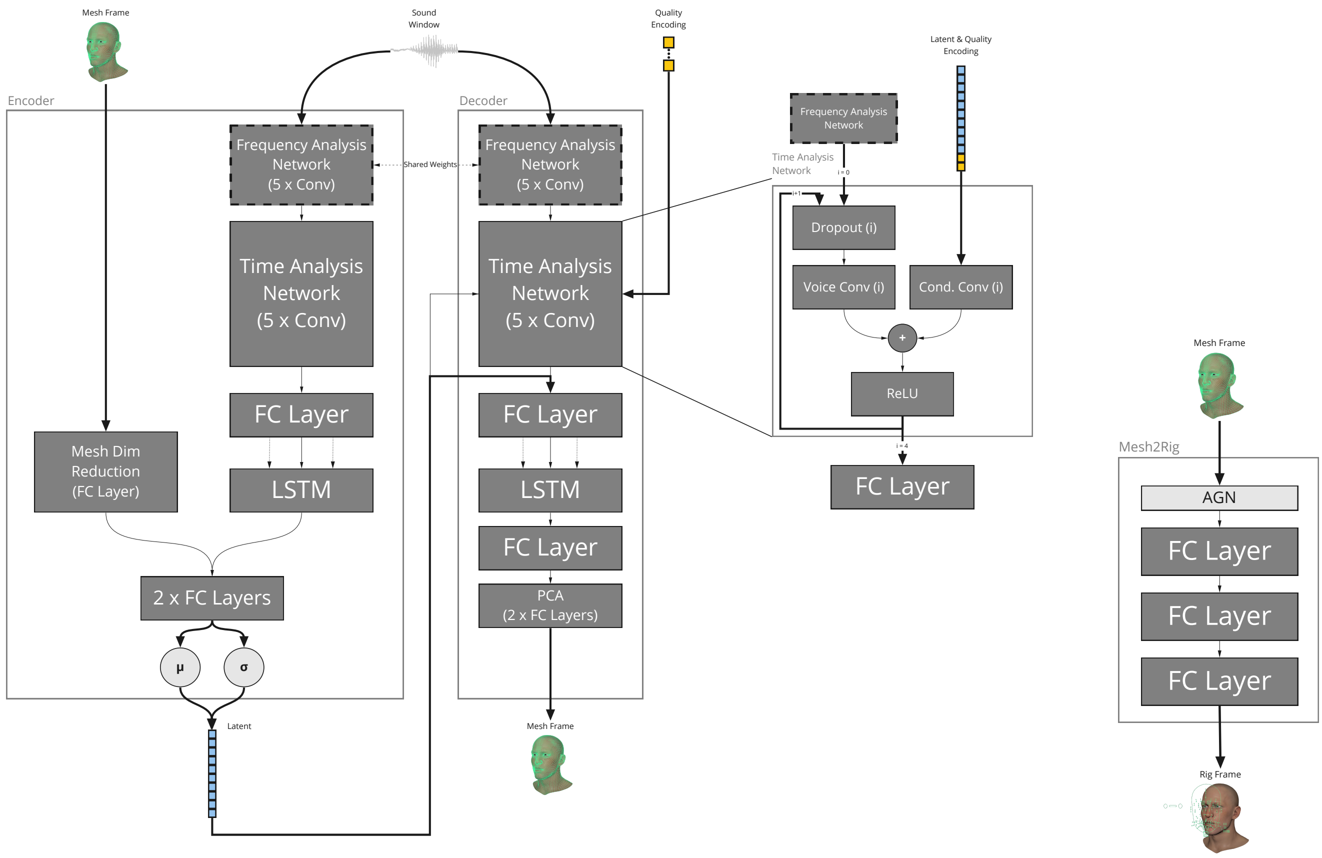
ACM SIGGRAPH / Eurographics Symposium on Computer Animation 2022
D. L. Michels and S. Pirk
(Guest Editors)

Volume 41 (2022), Number 8

Supplementary Material for *Voice2Face: Audio-driven Facial and Tongue Rig Animations with cVAEs*



(a) Facial Mesh Prediction Network

(b) Facial Mesh2Rig

Figure 1: *Detailed architecture of the Voice2Face mesh prediction module and the Mesh2Rig translation module*

1. Mesh Generation Architectural Details

The network architectures of the mesh generation cVAE model are shown in Table 1 and 2 (encoder E and decoder D respectively). We use the following notation for conciseness: kernel size (K), stride size (S), padding size (P).

We also use the following symbols to substitute implementation details: $B = 64$ sound samples or bins, $F = 39$ sound features, $K = 3$ sequence of windows used as context for the LSTM, $V = 21213$ number of vertex coordinates, $LSTM = 150$ LSTM output size, $MDR = 150$

Mesh Dimensionality Reduction output size, $TAN = 256$ Time Analysis Network output size, $Z = 10$ latent size, $Q = 2$ one-hot encoding size, $PCA = 49$ PCA basis.

Encoder

Alias	Dropout	Layer	Input/Output	Activation
Frequency Analysis Network (*)	0.1	Conv(K1x3, S1x2, P0x1)	$B{\times}F{\times}1 \rightarrow 64{\times}20{\times}72$	ReLU
	0.1	Conv(K1x3, S1x2, P0x1)	$64{\times}20{\times}72 \rightarrow 64{\times}10{\times}108$	ReLU
	0.1	Conv(K1x3, S1x2, P0x1)	$64{\times}10{\times}108 \rightarrow 64{\times}5{\times}162$	ReLU
	0.1	Conv(K1x3, S1x2, P0x1)	$64{\times}5{\times}162 \rightarrow 64{\times}3{\times}243$	ReLU
	0.1	Conv(K1x3, S1x1)	$64{\times}3{\times}243 \rightarrow 64{\times}1{\times}256$	ReLU
Time Analysis Network	0.1	Conv(K3x1, S2x1, P1x0)	$64{\times}1{\times}256 \rightarrow 32{\times}1{\times}256$	ReLU
	0.1	Conv(K3x1, S2x1, P1x0)	$32{\times}1{\times}256 \rightarrow 16{\times}1{\times}256$	ReLU
	0.1	Conv(K3x1, S2x1, P1x0)	$16{\times}1{\times}256 \rightarrow 8{\times}1{\times}256$	ReLU
	0.1	Conv(K3x1, S2x1, P1x0)	$8{\times}1{\times}256 \rightarrow 4{\times}1{\times}256$	ReLU
	0.1	Conv(K4x1, S1x1)	$4{\times}1{\times}256 \rightarrow 1{\times}1{\times}256$	ReLU
FC Layer	0	Linear	$256 \rightarrow 150$	ReLU
LSTM	0	LSTM	$K{\times}150 \rightarrow 150$	-
Mesh Dim Reduction	0	Linear	$V \rightarrow 150$	ReLU
2 x FC Layers	0	Linear	$LSTM + MDR \rightarrow 200$	ReLU
	0	Linear	$200 \rightarrow 2Z$	-

Table 1: *Detailed description of the architecture of the encoder in the mesh generation network. The Frequency Analysis Network is shared with the decoder. The alias Mesh Dim Reduction is part of a branch parallel to the LSTM as shown in Fig. (a).*

Decoder

Alias	Dropout	Layer	Input/Output	Activation
Frequency Analysis Network (*)	0.1	Conv(K1x3, S1x2, P0x1)	$64{\times}39{\times}1 \rightarrow 64{\times}20{\times}72$	ReLU
	0.1	Conv(K1x3, S1x2, P0x1)	$64{\times}20{\times}72 \rightarrow 64{\times}10{\times}108$	ReLU
	0.1	Conv(K1x3, S1x2, P0x1)	$64{\times}10{\times}108 \rightarrow 64{\times}5{\times}162$	ReLU
	0.1	Conv(K1x3, S1x2, P0x1)	$64{\times}5{\times}162 \rightarrow 64{\times}3{\times}243$	ReLU
	0.1	Conv(K1x3, S1x1)	$64{\times}3{\times}243 \rightarrow 64{\times}1{\times}256$	ReLU
Time Analysis Network	0.1	Conv(K3x1, S2x1, P1x0)	$64{\times}1{\times}256 \rightarrow 32{\times}1{\times}256$	-
	0	Conv(K1x1, S1x1)	$1{\times}1{\times}(Z+Q) \rightarrow 1{\times}1{\times}256$	-
	0	Add	$32{\times}1{\times}256 \rightarrow 32{\times}1{\times}256$	ReLU
	0.1	Conv(K3x1, S2x1, P1x0)	$32{\times}1{\times}256 \rightarrow 16{\times}1{\times}256$	-
	0	Conv(K1x1, S1x1)	$1{\times}1{\times}(Z+Q) \rightarrow 1{\times}1{\times}256$	-
	0	Add	$16{\times}1{\times}256 \rightarrow 16{\times}1{\times}256$	ReLU
	0.1	Conv(K3x1, S2x1, P1x0)	$16{\times}1{\times}256 \rightarrow 8{\times}1{\times}256$	-
	0	Conv(K1x1, S1x1)	$1{\times}1{\times}(Z+Q) \rightarrow 1{\times}1{\times}256$	-
	0	Add	$8{\times}1{\times}256 \rightarrow 8{\times}1{\times}256$	ReLU
	0.1	Conv(K3x1, S2x1, P1x0)	$8{\times}1{\times}256 \rightarrow 4{\times}1{\times}256$	-
	0	Conv(K1x1, S1x1)	$1{\times}1{\times}(Z+Q) \rightarrow 1{\times}1{\times}256$	-
	0	Add	$4{\times}1{\times}256 \rightarrow 4{\times}1{\times}256$	ReLU
	0.1	Conv(K4x1, S1x1)	$4{\times}1{\times}256 \rightarrow 1{\times}1{\times}256$	-
	0	Conv(K1x1, S1x1)	$1{\times}1{\times}(Z+Q) \rightarrow 1{\times}1{\times}256$	-
	0	Add	$1{\times}1{\times}256 \rightarrow 1{\times}1{\times}256$	ReLU
FC Layer	0	Linear	$TAN + Z \rightarrow 150$	ReLU
LSTM	0	LSTM	$K{\times}150 \rightarrow 150$	-
FC Layer	0	Linear	$150 \rightarrow 150$	ReLU
PCA	0	Linear	$150 \rightarrow PCA$	-
	0	Linear	$PCA \rightarrow V$	-

Table 2: *Detailed description of the architecture of the decoder in the mesh generation network. The Frequency Analysis Network is shared with the encoder. The additive layers (Add) sum the output of the two previous convolutions that process the abstract sound representation, and the concatenation of the latent and the quality condition, respectively, as shown in Fig. (a).*

2. Mesh2Rig Architectural Details

Similar to the mesh generation, the architecture of the mesh to rig mapping network is shown in Table 3, where $V = 21213$ is the number of vertex coordinates and $P = 78$ is the number of rig parameter controllers.

Mesh2Rig

Alias	Layer	Input/Output	Activation
AGN	GaussianNoise(μ=0, σ=0.3)	$V \rightarrow V$	-
FC Layer	Linear	$V \rightarrow 350$	ReLU
FC Layer	Linear	$350 \rightarrow 350$	ReLU
FC Layer	Linear	$350 \rightarrow P$	-

Table 3: *Detailed description of the Mesh2Rig network architecture. The Gaussian Noise layer samples noise during training, centered around zero with the empirical variance $\sigma = 0.3$ found in our experiments. The noise is added to the input signal before it's used as input to the first FC layer.*