Synthesizing Photorealistic Virtual Humans Through Cross-modal Disentanglement (Supplementary Material)

Siddarth Ravichandran, Ondřej Texler, Dimitar Dinev, Hyun Jae Kang
NEON, Samsung Research America
{siddarth.r, o.texler, dimitar.d, hyunjae.k}@samsung.com

Additional Results. In Fig. 4 we show additional results on a diverse set of identities, three females and two males, all of a different ethnicity and skin color. As demonstrated by the second row, our method can handle *floppy* hair. Also, it is able to faithfully preserve the facial hair, see the last row.

Mouth Decoder Ablation Study. Human faces have a great range of different frequencies of motion in different face regions. In order to accurately capture these various frequencies, we task the head and mouth decoders of our network to learn different regions of the face from the same latent space, offering us higher image quality and correctness for generating the head and mouth images respectively. We conduct an ablation study to show that without this 2-decoder arrangement and feature pasting, the quality of images generated by the head generator is inferior Fig. 1a compared to Fig. 1b.

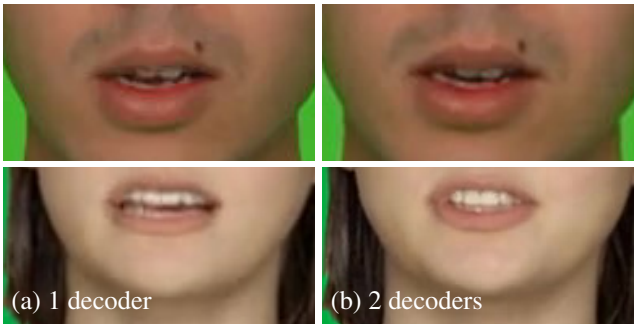

Figure 1. Image quality of the mouth region produced by the two decoder model (b) is superior compared to a single decoder model (a). Teeth in (a) do not have correct shape and appear blurry.

Alternative Audio Representations. Since visemes can be difficult and impractical to obtain, we tested our method with alternative audio representations. Our system trained using visemes Fig. 2a produces comparable results to one trained using wav2vec Fig. 2b. This is because both visemes and wav2vec audio representations exhibit similar properties: both are frame-level features with an adequate degree of voice and language independence. Please refer to the supplementary material for further discussion.

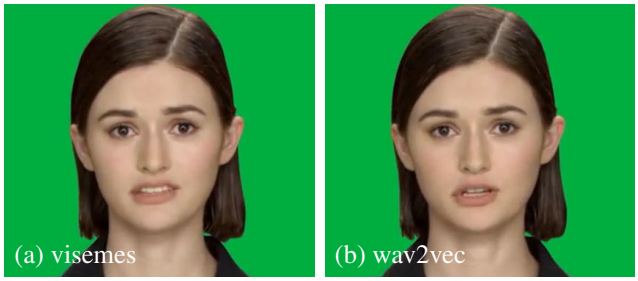

Figure 2. Our system is not limited to any particular audio representation and can work, for example, with visemes (a) and wav2vec (b). While the two images look slightly different even though they are supposed to pronounce the same sound, both are a correct representation of the sound 'ee'.

Representation of Input Data. Fig. 3 shows an illustration of the input features our system uses. The contour drawing is represented as a one channel image, the visemes are represented as a 1D vector.

(a) conotur drawing (b) visemes

Figure 3. Visualization of the input data. (a) contour drawing, (b) visualization of visemes.

Voice and Language Independence. Since visemes are

Layer	W	S	IP	OP	C_{in}	C_{out}	Output
(a) Keypoint Encoder							
Conv	3	2	1		3	8	$8 \times 256 \times 256$
Conv	3	2	1		8	16	$16 \times 128 \times 128$
Conv	3	2	1		16	32	$32 \times 64 \times 64$
Conv	3	2	1		32	64	$64 \times 32 \times 32$
(b) Audio Encoder							
Conv1d	1	1	0		6	1	1×34
MLP					34	1024	1024
MLP					1024	1024	1024
Reshape							$1 \times 32 \times 32$
(c) Fusion							
Cat							$65 \times 32 \times 32$
Conv	3	1	1		65	64	$64 \times 32 \times 32$
(d) Mouth Decoder							
TConv	3	2	1	1	64	32	$32 \times 64 \times 64$
TConv	3	2	1	1	32	16	$16 \times 128 \times 128$
TConv	3	2	1	1	16	8	$8 \times 256 \times 256$
TConv	3	2	1	1	8	8	$8 \times 512 \times 512$
Conv	3	1	1		8	8	$8 \times 512 \times 512$
Conv	7	1	3		8	3	$3 \times 512 \times 512$
(e) Face Decoder							
TConv	3	2	1	1	64	32	$32 \times 64 \times 64$
TConv	3	2	1	1	32	16	$16 \times 128 \times 128$
TConv	3	2	1	1	16	8	$8 \times 256 \times 256$
TConv	3	2	1	1	8	8	$8 \times 512 \times 512$
Conv	3	1	1		8	8	$8 \times 512 \times 512$
Conv	7	1	3		8	3	$3 \times 512 \times 512$

Table 1. Details of the architecture of neural network blocks used in our framework. **Conv** comprises of a Convolution2D layer, Instance Normalization, and a LeakyReLU(0.2) activation. **TConv** comprises of a TransposeConvolution2D layer, Instance Normalization, and LeakyReLU(0.2) activation. **W** is a kernel size, **S** is a stride, **IP** is an input padding, **OP** is an output padding, C_{in} is number of input channels, and C_{out} is number of output channels.

geometric shapes, they are largely independent of the voice characteristics. Because of this, our model, trained on a single voice, is able to produce good lip motion on a variety of voices, including natural voices from other people and synthetic voices from, e.g., text-to-speech. For the same reasons, visemes also provide some level of language independence. However, different languages can have very different phonetic distributions in common speech. When considering two languages that have similar distributions, such as English and Spanish, our English-trained model performs well on Spanish audio. However, when we use Korean audio with the same model, there is a degradation in the lip sync quality. These results are shown in the supplementary video.

Network Architecture Details. Our network consist of several building blocks. It comprises of two encoders, one for contour drawings Table 1a and one for audio Table 1b. Fusion block Table 1c to combine the two modalities. We also utilize two decoders, one for synthesizing the mouth region Table 1d and the other to supervise synthesis of the entire head region Table 1e.

Figure 4. Results of our method on five different identities. Zoom into the figure to better see lips and overall texture quality.