EmoVOCA: Speech-Driven Emotional 3D Talking Heads. Supplementary Materials

Federico Nocentini
University of Florence, Italy
federico.nocentini@unifi.it

Claudio Ferrari
University of Parma, Italy
claudio.ferrari@unipr.it

Stefano Berretti
University of Florence, Italy
stefano.berretti@unifi.it

In this supplementary material, we provide additional details and results that did not fit into the main paper.

1. DE-SD Implementation Details

Meshes from both VOCAset and Florence 4D are in FLAME topology, thus it was possible to define two identical encoders. Each encoder is constructed using a concatenation of five spiral convolution layers and a downsampling layer. The convolutional filters of the encoder have size [3, 16, 32, 64, 128], respectively for the five layers, and each layer is followed by an *elu* activation function. Embedded features are of dimension 64. The decoder is constructed similarly by concatenating five spiral convolution layers, where the convolutional filters are a mirror of those used in the encoder, each followed by an *elu* activation function and an upsampling layer. The decoder takes as input a feature of size 128 (after concatenation of f_i^e and f_j^t), and reconstructs the facial displacements S_i^{et}. The framework is trained for 100 epochs over each dataset using the Adam optimizer [2], with a learning rate of 10^{-4} and a minibatch size of 64.

2. Ablation study on feature combination

In Section 3 of the main paper, we described the DE-SD architecture, and stated that the embeddings f_i^e and f_j^t are combined through concatenation. We explored different alternatives for combining the embeddings. In Table 1, we report the outcome of two additional methods, *i.e*., *summation* and *multiplication*. These strategies were applied both during the training and inference phases. For instance, if the choice was to sum or multiply them during training, we performed the respective operation on the features from both encoders with themselves. During the inference phase, we summed or multiplied the features from both encoders. Our findings evidence that concatenating the features results in a better reconstruction error compared to using summation or multiplication. Another advantage of features concatenation is that during the inference phase, the decoder can access the features extracted from both encoders. In contrast, when using summation or multiplication, the decoder encounters a single set of features representing contributions of both encoders.

Table 1. Ablation on different feature combination strategies.

Baseline	**LVE** (mm) ↓	**UVE** (mm) ↓
DE-SD sum	0.732	0.690
DE-SD mult	0.765	0.732
DE-SD concat	**0.722**	**0.657**

The superior efficiency of the double encoder framework compared to the single encoder framework is demonstrated in Fig. 1. We applied the t-SNE [5] algorithm to features extracted from both frameworks, revealing that the DE-SD setup significantly improves the model's ability to differentiate between emotional facial expressions.

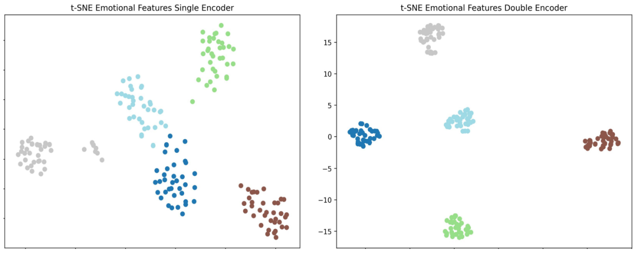

Figure 1. t-sne Single vs Double Encoder. When we apply the t-SNE algorithm to the features extracted from the encoders, it becomes evident that the DE setup is more effective in separating expressive faces.

3. Ethical Comments

As a concluding consideration, we acknowledge the ethical concerns of generating 3D facial animations. The production of synthetic narratives using generated 3D faces carries inherent risks and can lead to both intentional and unintentional consequences for individuals and society at large. We underscore the imperative of prioritizing a

Table 2. Primary expressions and their corresponding expressions in the Florence dataset.

Primary expression	**Expressions**
Anger, AR (6)	Angry1, Angry2, Fierce, Glare, Rage, Snarl
Fear, FR (6)	Afraid, Ashamed, Fear, Scream, Terrified, Worried
Sadness, SS (13)	Agony, Bereft, Ill, Mourning, Pain, Pouting, Pouty, Sad1, Sad2, Serious, Tired1, Tired2, Upset
Disgust, DT (9)	Arrogant, Bored, Contempt, Disgust, Displeased, Ignore, Irritated1, Irritated2, Unimpressed
Surprise, SE (11)	Awe, Confused, Ditzy, Drunk1, Frown, Hurt, Incredulous, Moody, Shock, Surprised, Suspicious
Anticipation, AN (4)	Cheeky, Concentrate, Confident, Cool
Trust, TT (6)	Desire, Drunk2, Flirting, Hot, Kissy, Wink
Joy, JY (15)	Amused, Dreamy, Excitement, Happy, Innocent, Laughing, Pleased, Sarcastic, Silly, Smile1, Smile2, Smile3, Smile4, Triumph, Zen

human-centered approach when developing and implementing such technology.

4. Datasets

The datasets have been split following state-of-the-art methods [1,3,4], with eight actors for training and two each for validation and testing. The EmoVOCA dataset utilized is just a subset of the possible versions. The expressions in the Florence dataset total 70 and are defined in Table 2.

References

[1] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3D facial animation with transformers. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 18770–18780, 2022. 2

[2] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 1

[3] Federico Nocentini, Claudio Ferrari, and Stefano Berretti. Learning landmarks motion from speech for speaker-agnostic 3D talking heads generation. In *Int. Conf. on Image Analysis and Processing (ICIAP)*, 2023. 2

[4] Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. Facediffuser: Speech-driven 3D facial animation synthesis using diffusion. In *ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG'23)*, 2023. 2

[5] Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008. 1