Supplementary
Style Transfer for Co-Speech Gesture Animation: A Multi-Speaker Conditional-Mixture Approach

Chaitanya Ahuja[1], Dong Won Lee[1], Yukiko I. Nakano[2], and Louis-Philippe Morency[1]

[1] Carnegie Mellon University, Pittsburgh, PA, USA
[2] {cahuja,dongwonl}@andrew.cmu.edu, morency@cs.cmu.edu
Seikei University, Musashino, Tokyo, Japan
y.nakano@st.seikei.ac.jp

1 PATS dataset

1.1 Speaker List

The list of speakers in the dataset are in Figure 1 as a dendrogram. This dendogram was created using text as the discriminating features. Speakers within the same cluster have a similar vocabulary. For the purposes of our experiments we use the speakers listed in Table 2 and 3.

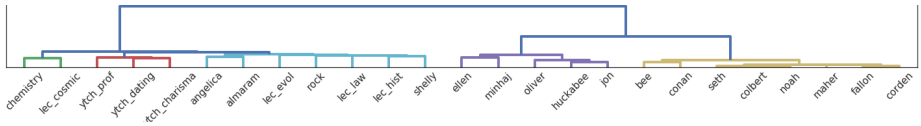


Fig. 1: List of speakers in the dataset as a dendrogram based on the content of the speech.

1.2 Attributes

We define 4 different attributes in Table 1 and select pairs of speakers that demonstrate visually striking differences with respect to those attributes.

Sitting/Standing	Gesture Frequency	Body Orientation	Primary Arm Func.
Sitting: Noah	**Low:** Seth	**Right:** Chemistry	**Right Arm:** lec_cosmic
Standing: Maher	**High:** Oliver	**Left:** Oliver	**Left Arm:** lec_cosmic

Table 1: Selection of speakers for attribute-level style modeling

2 Other Results, Discussions and Future Directions

These results complement Section 6 of the main paper with a few more observations, explorations and ablation studies. This is followed by potential future directions.

Attribute-Level Style Preservation Gesture generation for pairs of speakers shows improvements in PCK and F1 scores (see Table 2) following the trend of perceptual study in Figure 6 of the main paper.

Attributes	Speakers	Single-Speaker Models				Multi-Speaker Models					
		S2G		CMix-GAN		MUNIT		StAGE		Mix-StAGE	
		PCK	F1	PCK	F1	PCK	F1	PCK	F1	PCK	F1
Sitting/Standing	**Mean**	0.35	0.12	0.35	0.25	0.36	0.07	**0.42**	0.18	**0.42**	**0.25**
Sitting	Noah	0.45	0.11	0.45	0.28	0.34	0.09	0.44	0.14	0.44	0.26
Standing	Maher	0.25	0.13	0.24	0.22	0.22	0.07	0.28	0.26	0.26	0.25
Gesture Frequency	**Mean**	0.55	0.41	0.56	0.44	0.34	0.14	**0.58**	0.51	**0.58**	**0.53**
Low	Seth	0.56	0.50	0.58	0.54	0.22	0.02	0.58	0.54	0.59	0.57
High	Oliver	0.54	0.32	0.54	0.34	0.35	0.22	0.54	0.38	0.56	0.42
Body Orientation	**Mean**	0.39	0.14	**0.43**	0.25	0.14	0.05	0.40	**0.42**	0.40	**0.40**
Right	Chemistry	0.35	0.23	0.36	0.27	0.15	0.05	0.37	0.39	0.40	0.39
Left	lec_evol	0.44	0.05	0.50	0.23	0.28	0.34	0.50	0.44	0.49	0.46
Primary Arm Func.	**Mean**	0.43	0.12	0.43	0.33	0.35	0.02	0.59	0.30	**0.61**	**0.37**
Left Arm	lec_cosmic	0.41	0.08	0.41	0.28	0.32	0.06	0.60	0.27	0.62	0.36
Right Arm	lec_cosmic	0.45	0.18	0.45	0.38	0.44	0.12	0.58	0.31	0.60	0.35

Table 2: Objective metrics for attribute-level style preservation of single-speaker and multi-speaker models as indicated in the columns. Each row refers to the number of speakers the model was trained, with the average performance indicated at the top. The scores for common individual speakers are also indicated below alongside. For detailed results on other speakers please refer to the supplementary

Speaker-Level Style Preservation Complete numerical results for speaker-level style preservation (for Table 1 in the main paper) are listed in Table 3. The PCK and F1 scores of the individual speakers show the same trend as the average score for each model.

Impact of value of M on gesture generation We run an ablation study on the choice of M for the pose decoder. We report the average of PCK and F1 scores in Table **??** which were calculated for each speaker in single-speaker models. We find the the scores plateau with increasing values of M for single speaker models unlike multi-speaker models like Mix-StAGE .

No. of Speakers	Speaker	Single-Speaker Models				Multi-Speaker Models					
		S2G		CMix-GAN		MUNIT		StAGE		Mix-StAGE	
		PCK	F1	PCK	F1	PCK	F1	PCK	F1	PCK	F1
2	**Mean**	0.25	0.08	0.26	**0.27**	0.24	0.06	0.36	0.21	**0.34**	0.22
	Corden	0.30	0.05	0.32	0.21	0.25	0.06	0.36	0.21	0.34	0.22
	lec_cosmic	0.19	0.12	0.19	0.33	0.15	0.19	0.20	0.48	0.24	0.49
4	**Mean**	0.37	0.18	0.37	0.27	0.22	0.03	0.38	**0.34**	**0.39**	**0.35**
	Corden	0.30	0.05	0.32	0.21	0.24	0.07	0.35	0.27	0.35	0.30
	lec_cosmic	0.19	0.12	0.19	0.33	0.19	0.16	0.18	0.23	0.20	0.19
	ytch_prof	0.43	0.22	0.43	0.22	0.15	0.02	0.42	0.34	0.40	0.32
	Oliver	0.54	0.32	0.54	0.34	0.20	0.09	0.54	0.47	0.55	0.52
8	**Mean**	0.36	0.14	0.37	0.26	0.31	0.15	0.38	**0.32**	**0.40**	**0.33**
	Corden	0.30	0.05	0.32	0.21	0.23	0.03	0.32	0.28	0.36	0.27
	lec_cosmic	0.19	0.12	0.19	0.33	0.13	0.09	0.23	0.34	0.24	0.32
	ytch_prof	0.43	0.22	0.43	0.22	0.39	0.37	0.44	0.39	0.45	0.39
	Oliver	0.54	0.32	0.54	0.34	0.35	0.30	0.54	0.39	0.54	0.46
	Ellen	0.29	0.13	0.30	0.23	0.33	0.17	0.34	0.21	0.33	0.25
	Noah	0.45	0.11	0.45	0.28	0.40	0.23	0.44	0.24	0.44	0.27
	lec_evol	0.44	0.05	0.50	0.23	0.33	0.42	0.45	0.66	0.48	0.66
	Maher	0.25	0.13	0.24	0.22	0.23	0.17	0.25	0.25	0.25	0.25

Table 3: Objective metrics for speaker-level style preservation of single-speaker and multi-speaker models as indicated in the columns. Each row refers to the number of speakers the model was trained, with the average performance indicated at the top. The scores for individual speakers are also indicated below alongside. * refers to a Single-speaker Model

Single Speaker Models	Metrics	
	F1 ↑	PCK ↑
S2G	18.9	36.6
CMix-GAN ($M = 1$)	26.6	**37.9**
CMix-GAN ($M = 4$)	27.7	36.6
CMix-GAN ($M = 8$)	**28.0**	36.7
CMix-GAN ($M = 12$)	**27.8**	**37.0**

Table 4: Comparision of Mix-StAGE with different values of M over F1 and PCK. The results are reported as a mean over all speakers in PATS. We can see that the performance for single speaker models does not improve by increasing the number of modes M. This is unlike multi-speaker models, where the addition of sub-generators gives the model an edge over single-speaker models.

Exploring Style Control As a preliminary experiment, we modified the style vector to $[0.5, 0.5]$ in order to mix the styles of two speakers with different Primary Arm Functions. The generated gesture space in Figure 2 indicates that different speaker styles could be interpolated into a completely new style.

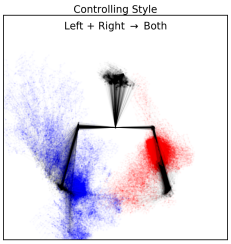


Fig. 2: Heat map of mixture of two styles: primary arm function of left and right mixed to give motion for both hands. Red represents the left hand and blue represents the right hand.

Future Directions Our efforts were aimed at modeling, disentangling and transferring gesture style under the assumption that the emotional state of the speaker does not affect the gestures. While this is reasonable for speakers in PATS, which are mostly scripted monologues, it may not be true in general hence motivating an interesting future direction. Another direction, that might induce diversity in the generated gestures, is the inclusion of verbal information (i.e. natural language). This may not be trivial in context of style transfer as the difference in vocabulary of different speakers could create an unwanted bias - some words might get associated with certain styles of gesturing.

3 Implementation Details

This section gives more detail about the exact architectures used for our model also described in Section 4 of the main paper.

3.1 Network Architectures

Figure 3 and 4 consists of the visual representation of the architectures used for our model Mix-StAGE . For the decoder $\bigoplus$ is a weighted sum as described in Equation (2) of the main paper. Every operation is a **1D-convolution** followed by a **Batch-Norm** and finally **ReLU**. Each convolution uses a kernal size of 3 and hop length of 1, except for cases where temporal dimension is downsampled where the kernal size is 4 and hop length is 2.

3.2 Training Details

We use Adam [1] to optimize the model with a exponentially decaying learning rate of 0.001. We train each model for 60000 iterations while check-pointing every 3000 iterations. Finally, we choose the best model based on loss on the development set. We use $\lambda_{id} = 0.1$ to prevent the style consistency loss from stealing focus while training the pose gesture generator.

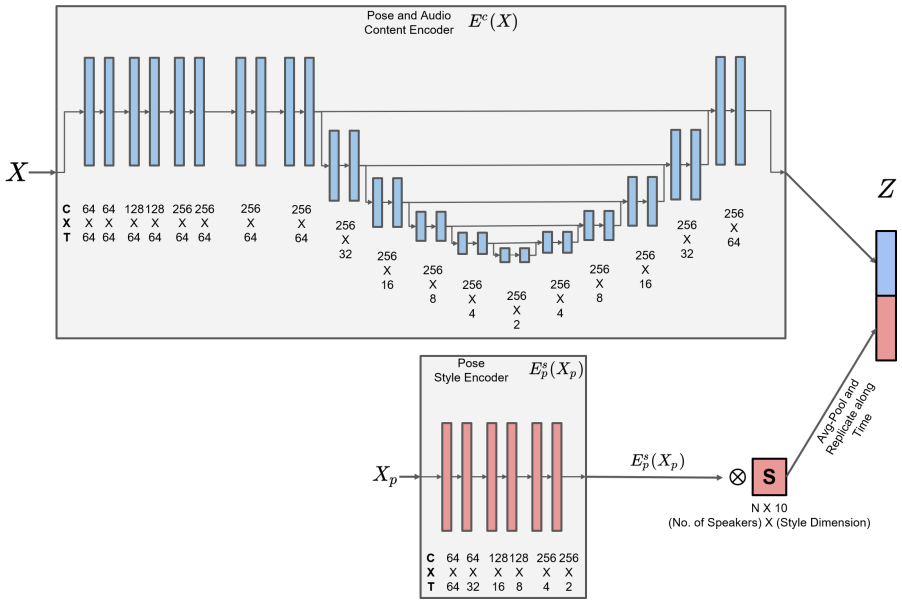


Fig. 3: Encoder Architecture

3.3 Human study experiments

We conducted our human studies on Amazon Mechanical Turk (or AMT), for which we used 100 random videos for each speaker which gave us 2400 pairs of comparisons per model for each study (out of 5). This is a significant number of comparisons and helps with reliability of the results. Each annotation task contained 20 videos and is performed by 3 different users; hence we had approximately (2400*3/20) = 360 participants.

To help filter unreliable annotators, we use two ground truth videos from the same speaker with the same style as control samples. If annotators tag these two videos as different styles, then we disregard this annotation set as unreliable.

Sample study for style transfer (other studies also follow similar method)
Two videos are shown to the user. One video is a ground truth(Speaker A Style A) and

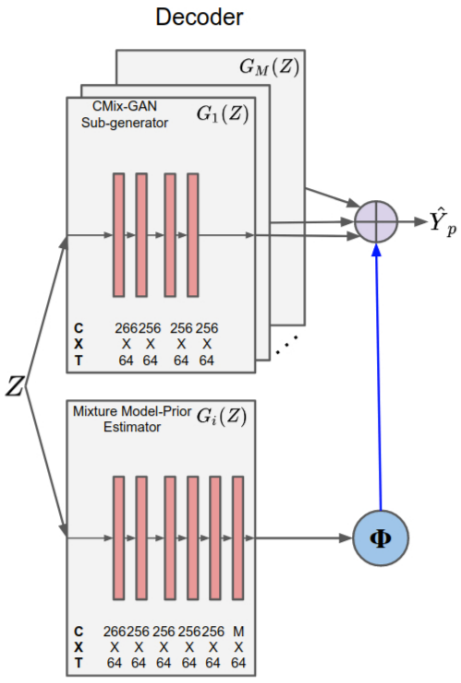


Fig. 4: Decoder Architecture

the other is generated by a model. The generated video could be either of (a) Speaker A Style A or (b) Speaker B Style A. We ask two questions to measure correctness of style transfer and naturalness:

1. Do the animations have different styles of gestures?
2. Which of the videos 1 or 2 has more Natural gestures with respect to the audio?

4 Videos: Style Transfer and Preservation

We refer the readers to `http://chahuja.com/mix-stage` for demo videos.

5 Video Frames: Style Preservation Qualitative Results

These results complement Figure 8 in the main paper. We plot some more animation figures generated by random audio samples in the test-set to provide some more samples for qualitative judgment in Figure 5.

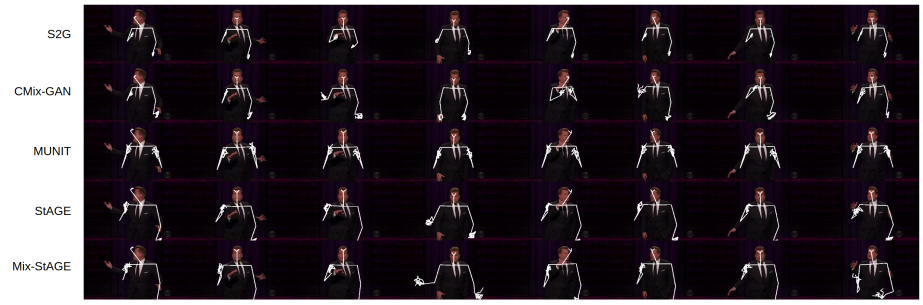
References

1. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 5

Fig. 5: Animation depicted as a series of frames for different speakers. The vertical axis is labeled as models and horizontal axis is time. The generated animation is superimposed over the ground truth video.

oliver

lec_cosmic



corden

ytch_prof