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

Chaitanya Ahuja[1 [0000−0003−4396−2050]], Dong Won Lee[1], Yukiko I. Nakano[2], and Louis-Philippe Morency[1 [0000−0001−6376−7696]]

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

Abstract. *How can we teach robots or virtual assistants to gesture naturally? Can we go further and adapt the gesturing style to follow a specific speaker?* Gestures that are naturally timed with corresponding speech during human communication are called co-speech gestures. A key challenge, called gesture style transfer, is to learn a model that generates these gestures for a speaking agent 'A' in the gesturing style of a target speaker 'B'. A secondary goal is to simultaneously learn to generate co-speech gestures for multiple speakers while remembering what is unique about each speaker. We call this challenge style preservation. In this paper, we propose a new model, named Mix-StAGE, which trains a single model for multiple speakers while learning unique style embeddings for each speaker's gestures in an end-to-end manner. A novelty of Mix-StAGE is to learn a mixture of generative models which allows for conditioning on the unique gesture style of each speaker. As Mix-StAGE disentangles style and content of gestures, gesturing styles for the same input speech can be altered by simply switching the style embeddings. Mix-StAGE also allows for style preservation when learning simultaneously from multiple speakers. We also introduce a new dataset, Pose-Audio-Transcript-Style (PATS), designed to study gesture generation and style transfer. Our proposed Mix-StAGE model significantly outperforms the previous state-of-the-art approach for gesture generation and provides a path towards performing gesture style transfer across multiple speakers. Link to code, data and videos: http://chahuja.com/mix-stage.

Keywords: Gesture animation · Style transfer · Co-speech gestures

1 Introduction

Nonverbal behaviours such as body posture, hand gestures and head nods play a crucial role in human communication [55,41]. Pointing at different objects, moving hands up-down in emphasis, and describing the outline of a shape are some of the many gestures that co-occur with the verbal and vocal content of communication. These are known as co-speech gestures [38,27]. When creating new robots or embodied virtual assistants designed to communicate with humans, it is important to generate *naturalistic* looking gestures that are meaningful with the speech [6]. Some recent works have proposed speaker-specific gesture generation models [18,13,11,16] that are both trained

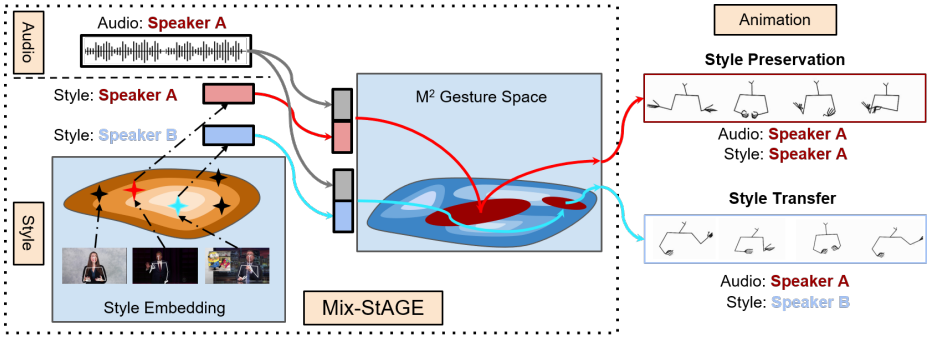


Fig. 1: Overview of co-speech gesture generation and gesture style transfer/preservation task. The models learns a style embedding for each speaker, which can be be mapped to a gesture space with either the same speaker's audio to generate style preserved gestures or a different speaker's audio to generate style transferred gestures.

and tested on the same speaker. The intuition behind this prior work is that co-speech gestures are idiosyncratic [57,38]. There is an unmet need to learn generative models that are able to learn to generate gestures simultaneously from multiple speakers (→ in Figure 1) while at the same time remembering what is unique for each speaker's gesture style. These models should not simply remember the "average" speaker. A bigger technical challenge is to be able to transfer gesturing style of speaker 'B' to speaker 'A' (→ in Figure 1).

The gesturing style can defined along two dimensions which is a result of (a) the speaker's idiosyncrasy (or speaker-level style), and (b) due to some more general attributes such as standing versus sitting, or the body orientation such as left versus right (or attribute-level style). For both gesture style types, the generation model needs to be able to learn the diversity and expressivity [42,8] present in the gesture space, within and amongst speakers. The gesture distribution is likely to have multiple modes, some of them shared among speakers and some distinct to a speaker's prototypical gestures.

In this paper, we introduce the Mixture-Model guided Style and Audio for Gesture Generation (or Mix-StAGE) approach which trains a single model for multiple speakers while learning unique style embeddings for each speaker's gestures in an end-to-end manner (see Figure 1). We use this model to perform two tasks for gesture generation conditioned on the input audio signal, (1) **style preservation** which ensures that while learning from multiple speakers we are still able to preserve unique gesturing styles of each speaker, and (2) **style transfer** where generated gestures are from a new style that was not the same as the source of the speech. A novelty of Mix-StAGE is to learn a mixture of generative models which allows for conditioning on the unique gesture style of each speaker. Our experiments study the impact of multiple speakers on both style transfer and preservation. Our study focuses on the non-verbal components of speech asking the research question if we can predict gestures without explicitly modeling verbal signals. We also introduce a new dataset, Pose-Audio-Transcript-Style (PATS), designed to study gesture generation and style transfer.

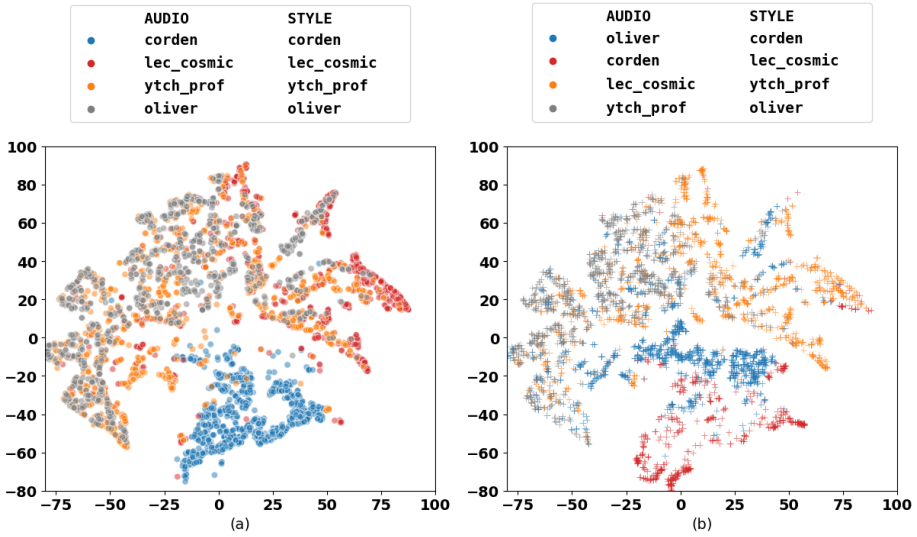


Fig. 2: t-SNE[37] representation of the Multi-mode Multimodal Gesture Space (Section 4.1). Each color represents a style, which is fixed for both plots. The plot on the left visualizes the gesture space generated from the audio content and style of the same speaker. The plot on the right shows the generated gesture space where the audio content and style are not from the same speaker. It can be observed that a similar gesture space is occupied by each speaker's style even when the audio content is not of their own.

2 Related Work

Speech driven Gesture Generation: For prosody-driven head motion generation [49] and body motion generation [32,31], Hidden Markov Models were used to predict a sequence of frames. Chiu & Marsella [12] proposed a two-step process: predicting gesture labels from speech signal using conditional random fields (CRFs) and converting the label to gesture motion using Gaussian process latent variable models (GPLVMs). More recently, an LSTM network was applied to MFCC features extracted from speech to predict a sequence of frames for gestures [22] and body motions [50,1]. Generative adversarial networks (GAN) were used to generate head motions [47] and body motions[16]. Gestures driven by an audio signal[18] is the closest approach to our task of style preservation but it uses models trained on single speakers unlike our multi-speaker models.

Disentanglement and Transfer of Style : Style extraction and transfer have been studied in context of image artistic style [17,26], factorizing foreground and background in videos[15,54], disentanglement in speech [56,9,20]. These approaches were extended to translation between properties of style such as map edges and real photos using paired samples [25]. Paired data limits the variety of attributes of source and target, which encouraged unsupervised domain translation for images [58,59] and videos[7]. Style

was disentangled from content using a shared latent space[33], a cycle consistency loss [58] and contrastive learning [39]. Cycle consistency losses were shown to limit diversity in the generated outputs as opposed to a weak consistency loss [24] and shared content space [29]. Cycle consistency in cross-domain translation assumes reversibility (i.e. domain A can be translated to domain B and vice-versa). These assumptions are violated in cross-modal translation [36] and style control [56] tasks where information in modality B (e.g. pose) is a subset of that in modality B (e.g. audio). Style transfer for pose has been studied in context of generating dance moves based on the content of the audio [30] or walking styles [52]. Generated dance moves are conditioned on both the style and content of the audio (i.e. kind of music like ballet or hip-hop), unlike co-speech gesture generation which requires only the content and not the style of the audio (i.e. speaker specific style like identity or fundamental frequency). Co-speech gesture styles have been studied in context of speaker personalities [40], but requires a long annotation process to create a profile for each speaker. To our knowledge, this is the first fully data-driven approach that learns gesture style transfer for multiple speakers in a co-speech gesture generation setting.

3 Stylized Co-Speech Gesture Animation

We define the problem of stylized co-speech gesture animation with two main goals, (1) generation of an animation which represents the gestures that would co-occur with the the spoken segment and (2) modification of the style of these gestures. Figure 1 shows the first goal (→) exemplified with the style preservation scenario, while the second goal (→) exemplifies with the style transfer scenario.

Formally, given a sequence of T audio frames $\mathbf{X}_a \sim F_a$ and i^{th} speaker's style $S(i)$, the goal is to predict a sequence of T frames of 2-D poses $\mathbf{Y}_p \sim F_p$. Here F_a and F_p are the marginal distributions of the content of input audio and style of output pose sequences. To control pose generation by both style and audio, we learn a joint distribution over pose, audio and style $F_{p,a,s}$ which can be broken down into 3 parts

$$F_{p,a,s} = F_{p|\Phi} F_{\Phi|a,s} \cdot F_s \cdot F_a \tag{1}$$

where $F_{\Phi|a,s}$ is the distribution of the gesture space Φ conditioned on the audio and style of pose (Figure 1). We discuss the modelling of $F_{p|\Phi} F_{\Phi|a,s}$, F_a, and F_s in Section 4.1, 4.2 and 4.3 respectively.

4 Mix-StAGE: Mixture-Model guided Style and Audio for Gesture Generation

Figure 3 shows an overview of our Mix-StAGE model, including the training inference pathways. A first component of our Mix-StAGE model is the audio encoder E_a^c, which takes as input the spoken audio X_a. During training, we also have the pose sequence of the speaker Y_p. This pose sequence is decomposed into content and style, with two specialized encoders E_p^c and E_p^s. During training, the style for the pose sequence can either be concatenated with the audio or the pose content.

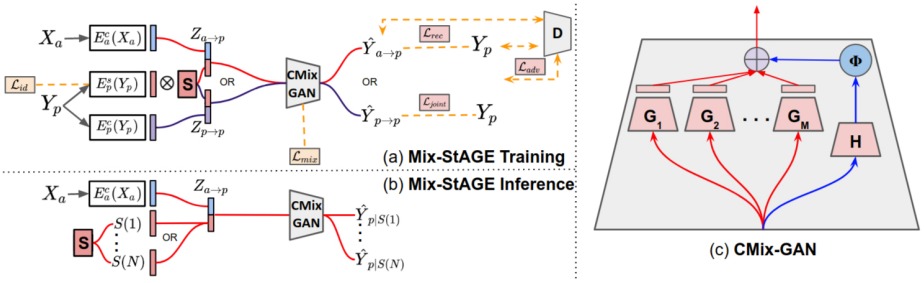


Fig. 3: (a) Overview of the proposed model Mix-StAGE in training mode, where audio X_a and pose Y_p are fed as inputs to learn a style embedding and concurrently generate a gesture animation. **S** represents the style matrix, which is multiplied with a separately encoded pose. $\bigotimes$ represents argmax for style ID followed by matrix multiplication. Discriminator D is used for adversarial training. All the loss functions are represented with dashed lines. (b) Mix-StAGE in inference mode, where any speaker's style embedding can be used on an input audio X_a to generate gesture style-transferred or style-preserved animations (c) CMix-GAN generator: a visual representation of the conditional Mix-GAN model, where the $\bigoplus$ represents a weighted sum of the model priors Φ with the generated outputs by the sub-generators.

The pose sequences for multiple speakers are represented as a distribution with multiple modes [21]. To decode from this multi-mode multimodal gesture space, we use a common generator G with multiple sub-generators (or CMix-GAN) conditioned on input audio and style to decode both these embeddings to output pose $\mathbf{Y}_p$.

Our loss function comprises of a mode-regularization loss (Section 4.1) to ensure that audio and style embedding can sample from the appropriate marginal distribution of poses, a joint loss (Section 4.2) to ensure latent distribution matching for content in a cross-modal translation task, a style consistency loss (Section 4.3) to ensure that the correct style is being generated and an adversarial loss (Section 4.4) that matches the generated pose distribution to the target pose distribution.

4.1 M^{2}GS: Multi-mode Multimodal Gesture Space

Humans perform different styles of gestures, where each style consists of different kinds of gestures (i.e beat, metaphorical, emblematic, iconic and so on)[38]. Learning pose generators for multiple speakers, each with their own style of gestures, presents a distribution with multiple modes. These gestures have a tendency of switching from one mode to the other over time, which depends on style embeddings and content of the audio.

To prevent mode collapse [4] we propose the use of mixture-model guided sub-generators [21,5,23], each learning a different mode of M^2 gesture space $F_{\Phi|a,s}$.

$$\hat{\mathbf{Y}}_p = \sum_{m=1}^{M} \phi_m G_m(\mathbf{Z}) = G(\mathbf{Z}) \tag{2}$$

where $\mathbf{Z} \in \{\mathbf{Z}_{a\to p}, \mathbf{Z}_{p\to p}\}$ are cross-modal and self-modal latent spaces respectively. They are defined as $\mathbf{Z}_{a\to p} = \left[E^c_a(\mathbf{X}_a), E^s_p(\mathbf{Y}_p) \bigotimes \mathbf{S}\right]$ and $\mathbf{Z}_{p\to p} = \left[E^c_p(\mathbf{Y}_p), E^s_p(\mathbf{Y}_p) \bigotimes \mathbf{S}\right]$ where $\mathbf{S}$ is the style embedding matrix (See Section 4.3) and $\bigotimes$ is argmax for style ID followed by matrix multiplication. Pose sequence $\hat{Y}_p \sim F_{p|\Phi}F_{\Phi|a,s}$ represents the pose probability distribution conditioned on audio and style. $G_m \sim F^m_{p|a,s}\ \forall m \in [1, 2, \dots M]$ are sub-generator functions with corresponding mixture-model priors $\Phi = \{\phi_1, \phi_2, \dots \phi_M\}$. These mixture model priors represent the M^2 gesture space and are estimated at inference time conditioned on the input audio and style.

Estimating Mixture Model Priors: During training, we partition poses Y_p into M clusters using an unsupervised approach, Lloyd's algorithm [35]. While other unsupervised clustering methods [43] can also be used at this stage, we choose Lloyd's algorithm for its simplicity and speed. Each of these clusters represent samples from probability distributions $\{F^1_{p|a,s}, F^2_{p|a,s}, \dots F^M_{p|a,s}\}$. If a sample belongs to the m^{th} cluster, $\phi_m = 1$, otherwise $\phi_m = 0$, making Φ a sequence of one-hot vectors. While training the generator G with loss function $\mathcal{L}_{\text{rec}}$, if a sample belongs to the distribution $F^m_{p|a,s}$, only parameters of sub-generator G_m are updated. Hence, each sub-generator learns different components of the true distribution, which are combined using Equation 2 to give the generated pose.

At inference time, we do not have the true values of mixture-model priors Φ. As mixture model priors modulate based on the style of the speaker and audio content at any given moment, we jointly learn a classification network $H \sim F_{\Phi|a,s}$ to estimate values of Φ in form of a mode regularization loss function

$$\mathcal{L}_{mix} = \mathbb{E}_{\Phi,\mathbf{Z}}\text{CCE}(\Phi, H(\mathbf{Z})) \tag{3}$$

where CCE is categorical cross-entropy.

4.2 Joint Space of Style, Audio and Pose

A set of marginal distributions F_a and F_s are learnt by our content encoders E^c_a and E^c_p, which together define the joint distribution of the generated poses: $F_{p,a,s}$. Since both cross-modal $\mathbf{Z}_{a\to p}$ and self-modal $\mathbf{Z}_{p\to p}$ latent spaces are designed to represent the same underlying content distribution, they should be consistent with each other. Using the same generator G for decoding both of these embeddings[34] yields content invariant generator. We enforce a reconstruction and joint loss [2] which encourages a reduction in distance between $\mathbf{Z}_{a\to p}$ and $\mathbf{Z}_{p\to p}$. As cross-modal translation is not reversible for this task (i.e. audio signal cannot be generated with pose input), a bi-directional reconstruction loss [29] for latent distribution matching cannot be directly used. This joint loss achieves the same goal of latent distribution matching in a uni-modal translation task [24,45,46] but for a cross-modal translation task.

$$\mathcal{L}_{joint} = \mathbb{E}_{\mathbf{Y}_p}\|\mathbf{Y}_p - G(\mathbf{Z}_{p\to p})\|_1 \tag{4}$$

$$\mathcal{L}_{rec} = \mathbb{E}_{\mathbf{Y}_p,\mathbf{X}_a}\|\mathbf{Y}_p - G(\mathbf{Z}_{a\to p})\|_1 \tag{5}$$

4.3 Style Embedding

We represent style as a collection of embeddings $S(i) \in \mathbf{S} \sim F_s$, where $S(i)$ is the style of the i^{th} speaker in the style matrix $\mathbf{S}$. Style space and embeddings are conceptually similar to the GST (Global Style Token) layer [56] which decomposes the audio embedding space into a set of basis vectors or style tokens, but only one out of the two modalities in the stylized audio generation task [56,36] have both style and content. In our case, both audio and pose have style and content. To ensure that generator G is attending only to style of pose while ignoring style of the audio, a style consistency loss is enforced on input $\mathbf{Y}_p$ and generated $\hat{\mathbf{Y}}_p$.

$$\mathcal{L}_{id} = \mathbb{E}_{Y \in \{\mathbf{Y}_p, \hat{\mathbf{Y}}_p\}} \text{CCE}\left(\text{Softmax}\left(E_p^s(Y)\right), \mathbf{ID}\right) \tag{6}$$

where **ID** is a one-hot vector denoting the speaker level style.

4.4 Total Loss with Adversarial Training

To alleviate the challenge of overly smooth generation caused by L1 reconstruction and joint losses in Equation 4,5, we use the generated pose sequence $\hat{\mathbf{Y}}^p$ as a signal for the adversarial discriminator D [18]. The discriminator tries to classify the true pose $\mathbf{Y}^p$ from the generated pose $\hat{\mathbf{Y}}^p$, while the generator learns to fool the discriminator by generating realistic poses. This adversarial loss[19] is written as:

$$\mathcal{L}_{adv} = \mathbb{E}_{\mathbf{Y}_p} \log D\left(\mathbf{Y}_p\right) + \mathbb{E}_{\mathbf{X}_a, \mathbf{Y}_p} \log\left(1 - D(G\left(\left[E_a^c(\mathbf{X}_a), E_p^s(\mathbf{X}_p)\right]\right))\right) \tag{7}$$

The model is jointly trained to optimize the overall loss function:

$$\max_{D} \min_{E_a^c, E_p^c, E_p^s, G} \mathcal{L}_{mix} + \mathcal{L}_{joint} + \mathcal{L}_{rec} + \lambda_{id}\mathcal{L}_{id} + \mathcal{L}_{adv} \tag{8}$$

where λ_{id} controls the weight of the style consistency loss term.

4.5 Network Architectures

Our proposed approach can work with any temporal network, giving it the flexibility of incorporating domain dependent or pre-trained temporal models.

In our experiments we use a Temporal Convolution Network (TCN) module for both content and style encoders. The style space is a matrix $\mathbf{S} \in \mathbb{R}^{N \times D}$ where N is the number of speakers and D is the length of the style embeddings. The generator $G(.)$ consists of a 1D version of U-Net [44,18] followed by M TCNs as sub-generator functions. The discriminator is also a TCN module with lower capacity than the generators. A more detailed architecture can be found in the supplementary.

5 Experiments

Our experiments are divided into 2 sections, (1) **Style Preservation:** Generating co-speech gestures for multiple speakers with their own individualistic style, (2) **Style**

No. of Speakers	Speaker	Single-Speaker Models				Multi-Speaker Models					
		S2G[18]		CMix-GAN		MUNIT[24]		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	0.24
	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	0.05	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
8	**Mean**	0.36	0.14	0.37	0.26	0.31	0.21	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

Table 1: **Style Preservation**: Objective metrics for pose generation 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. Bold numbers indicate $p < 0.1$ in a bootstrapped two sided t-test.

Transfer: Generating co-speech gestures with content (or audio) of a speaker and gesture style of another speaker. Additionally, style transfer can be speaker-level as well as attribute-level. We choose visually distinguishable attribute-level styles: (1) body orientation, (2) gesture frequency, (3) primary arm function and (4) sitting/standing posture. We start by describing the baseline models followed by the evaluation metrics, which we will use to compare our model. We end this section with the description of our proposed dataset.

5.1 Baseline Models

Single-Speaker Models : These models are not designed to perform style transfer and hence are not included for those experiments.

- **Speech2Gesture [18]**: The closest work to co-speech gesture generation is one that only generates individualistic styles. We use the pre-trained models available from their code-base to render the videos. For rest of the speakers in PATS, we replicate their model, hyper-parameters and train speaker specific models.
- **CMix-GAN** (variant of our model): As an ablation, we remove the style embedding module and style consistency losses from our model Mix-StAGE . Hence, a separate model is required to be trained for each speaker for style preservation experiments.

Multi-speaker Models

- **MUNIT [24]**: The closest work to our style-transfer task is MUNIT which takes multiple domains of images (i.e. uni-modal). We modify the encoders and decoders to domain specific architectures (i.e. 1D convolutions for audio instead of 2D convolutions for images) while retaining the loss functions.
- **StAGE** (variant of our model): As an ablation, we fix the number of sub-generators in our model Mix-StAGE to one. This is equivalent to setting $M = 1$ in equation 2.

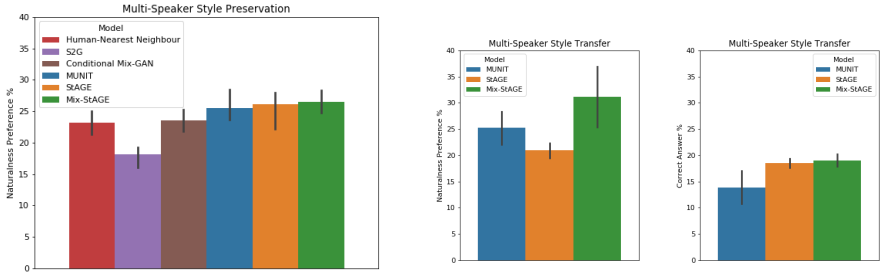
5.2 Evaluation Metrics

Human Perceptual Study: We conduct a human perceptual study on Amazon Mechanical Turk (AMT) for co-speech gesture generation (or style preservation) and style transfer (speaker-level and attribute-level) and measure preferences in two aspects of the generated animations, (1) **naturalness**, and (2) **style transfer correctness** for animation generation with content (i.e. audio) of speaker A and style of speaker B. We show a pair of videos with skeletal animations to the annotators. One of the animations is from the ground-truth set, while the other is generated using our proposed model. The generated animation could either have the same style or a different style as the original speaker. With unlimited time, the annotator has to answer two questions, (1) Which of the videos has more natural gestures? and (2) Do these videos have the same attribute-level style (or speaker-level style)? The first question is a real vs. fake perceptual study against the ground truth, while the second question measures how often the algorithm is able to visually preserve or transfer style (attribute or individual level). We run this study for randomly selected 100 pairs of videos from the held-out set. .

Probability of Correct Keypoints (PCK): To measure the accuracy of the gesture generation, PCK [3,51] is used to evaluate all models. PCK values are averaged over $\alpha = 0.1, 0.2$ as suggested in [18].

Mode Classification F1: Correctness of shape of a gesture can be quantified by measuring the number of times the model has sampled from the correct mode of the pose distribution. Formally, we use the true ($\mathbf{Y}_p$) and generated ($\hat{\mathbf{Y}}^p$) pose to find the closest cluster $\hat{m}$ and m respectively. If $m = \hat{m}$, the generated pose was sampled from the correct mode. F1 score of this M-class classification problem is defined as Mode Classification F1, or simply F1.

Inception Score (IS): Generated pose sequences with the audio of speaker A and style of speaker B does not have a ground truth reference. To quantitatively measure the correctness and diversity of generated pose sequence we use the inception score [48]. For generative tasks such as image generation, this metric has been used with a pre-trained classification network such as Inception Model [53]. In our case, the generated samples are not images, but a set of 2D keypoints. Hence, we train a network which classifies a sequence of poses to its corresponding speaker which estimates the conditional likelihood to calculate IS scores.



(a) Style Preservation Naturalness

(b) Style Transfer Naturalness

(c) Style Transfer Correctness

Fig. 4: Perceptual Study for speaker-level style preservation in (a) and speaker level style transfer in (b), (c). We have naturalness preference for both style transfer and preservation, and style transfer correctness scores for style transfer. Higher is better. Error bars calculated for $p < 0.1$ using a bootstrapped two sided t-test.

5.3 Pose-Audio-Transcript-Style (PATS) dataset 

Gesture styles, which may be defined by attributes such as type, frequency, orientation of the body, is representative of the idiosyncrasies of the speaker [14]. We create a new dataset, Pose-Audio-Transcript-Style (PATS), to study various styles of gestures for a large number of speakers in diverse settings.

PATS contains pose sequences aligned with corresponding audio signals and transcripts[3] for 25 speakers (including 10 speakers from [18]) to offer a total of 251 hours of data, with a mean of 10.7 seconds and a standard deviation of 13.5 seconds per interval. The demographics of the speakers include 15 talk show hosts, 5 lecturers, 3 YouTubers, and 2 televangelists.

Each speaker's pose is represented via skeletal keypoints collected via OpenPose [10] similar to [18]. It consists of of 52 coordinates of an individual's major joints for each frame at 15 frames per second, which we rescale by holding the length of each individual's shoulder constant. This prevents the model from encoding limb length in the style embeddings. Following prior work [28,18], we represent audio features as mel spetrograms, which is a rich input representation shown to be useful for gesture generation.

6 Results and Discussion

We group our results and discussions in (1) a first set of experiments studying style preservation (when output gesture styles are the same the original speaker) and (2) a second set of experiments studying transfer of gesture styles.

[3] While transcripts are a part of this dataset, they are ignored for the purposed of this work.

Model	Number of Speakers			Attributes			
	2 Speakers	4 Speakers	8 Speakers	Sitting vs Standing	Gesture Frequency	Body Orientation	Primary Arm Func.
MUNIT [24]	1.11	1.90	2.06	1.10	2.49	1.05	3.32
StAGE	2.17	**2.85**	3.89	1.68	4.38	**6.81**	3.14
Mix-StAGE	**2.61**	**2.85**	**4.48**	**3.08**	**4.50**	6.69	**3.32**

Table 2: **Style Transfer**: Inception scores for style transfer on multi-speaker models (indicated in each row). Columns on the left refer to the speaker-level style transfer task while those on the right refer to the specific attribute-level style task. Bold numbers indicate $p < 0.1$ in a bootstrapped two sided t-test.

6.1 Gesture Animation and Style Preservation

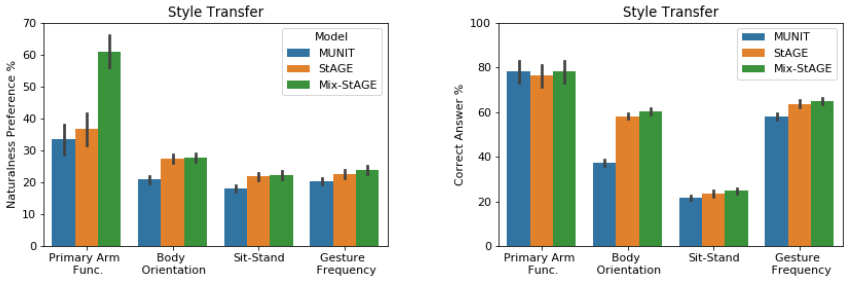
To understand the impact of adding more speakers, we select a random sample of 8 speakers for the largest 8-speaker multi-speaker model, and train smaller 4-speaker and 2-speaker models where the speakers trained are always a subset of the speakers that were trained in a larger model. This allows to compare the performance on the same two initial speakers which are '*Corden*' and '*lec_cosmic*' in our case[4]. We also compare with single-speaker models trained and tested on one speaker at a time.

Impact of training with Multiple Speakers Results from Table 1 show that multi-speaker models outperform single-speaker models especially for pose accuracy (i.e. PCK), shape and timing (i.e. F1). We find that increasing the number of speakers could sometimes reduce the performance of individual speakers but the overall performance generally shows improvement.

Comparison with previous baselines To compare with prior baselines, we focus first on the subjective evaluation shown in Figure 4a, since it is arguably the most important metric. The results show consistent improvements on the naturalness rating for our proposed model Mix-StAGE and also our single-speaker variant CMix-GAN over the previous state of the art approach S2G [18]. We also observe that multi-speaker models perform better than single speaker-models. In Table 1, we show similar quantitative improvements of Mix-StAGE and CMix-GAN over S2G for both PCK and F1 scores.

Impact of Multiple Generators for Decoding Mix-StAGE's gesture space models multiple modes, as seen in Figure 2. Its importance is shown in Table 1 where models with single generators as the decoder (i.e. S2G, MUNIT and StAGE) showed lower F1 scores, most likely due to mode collapse while training. Multiple generators in CMix-GAN and Mix-StAGE boost F1 scores as compared to other models in the single-speaker and multi-speaker regimes respectively. A similar trend was observed in the perceptual study in Figure 4.

[4] The complete set of speakers used in our experiments are listed in the supplementary.



(a) Naturalness Preference (b) Style Transfer Correctness

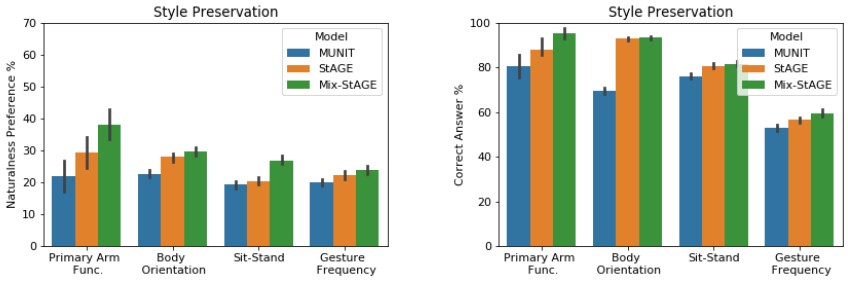
Fig. 5: A visualization of the perceptual human study for attribute-level style transfer with (a) naturalness preference, and (b) style transfer correctness scores for the generated animations for a different style than the speaker. Higher is better. Error bars calculated for $p < 0.1$ using a bootstrapped two sided t-test.

We also study the impact of the number of generators (hyperparameter M) in our Mix-StAGE model. While for small number of speakers (i.e. 2 speakers) a single generator is good enough, the positive effect of multiple generators can be observed as the number of speakers increase (see Table 1). We also vary $M \in \{1, 2, 4, 8, 12\}$ and observe that improvements seem to plateau at $M = 8$ with only marginal improvements for larger number of sub-generators. For the ablation study we refer the readers to the supplementary.

Attribute-level Style Preservation in Multi-Speaker Models We also study style preservation for attributes in Section 5 as a perceptual study in Figure 6. We observe that humans deem animations generated by Mix-StAGE significantly more natural in most cases. High scores ranging 60-90% for style preservation correctness, with Mix-StAGE outperforming others, are observed for pairs of speakers in Figure 6b. This indicates that style preservation may be a relatively easy task as compared to style transfer for multi-speaker models. With this, we now shift our focus to style transfer.

6.2 Style Transfer

Speaker-level Style Transfer To study our capability to transfer style of a specific speaker to a new speaker, we will compare the gesture spaces between the original speakers and the transferred speakers. Figure 2a shows that each original speaker occupies different regions in the M^2 gesture space. Using our Mix-StAGE model to transfer style, we can see the new gesture space in Figure 2b. For the transferred speakers the 2 spaces look quite similar. For instance, '*Corden*' style (a speaker in our dataset) is represented by the color blue in Figure 2a and occupies the lower region of the gesture space. When Mix-StAGE generates co-speech gestures using audio of '*Oliver*' and the style of '*Corden*', it occupies a subset of '*Corden's*' region in the gesture space, also represented by blue in Figure 2b. We see a similar trend for styles of '*Oliver*' and



(a) Naturalness Preference (b) Style Preservation Correctness

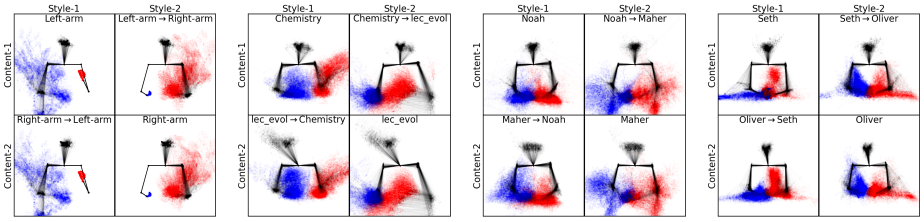
Fig. 6: A visualization of the perceptual human study for attribute-level style preservation with (a) naturalness preference, and (b) style preservation correctness scores for the generated animations for the same style as the speaker. Higher is better. Error bars calculated for $p < 0.1$ using a bootstrapped two sided t-test.

'*ytch_prof*'. This is an indication of a successful style transfer across different speakers. We note the lack of clean separation in the gesture space among different styles as there could common gestures across multiple speakers.

For the perceptual study, we want to know if humans can distinguish the generated speaker styles. For this, we show human annotators two videos: a ground truth video in a specific style, and a generated video which is either from the style of the same speaker or a different speaker. Annotators have to decide if this is the same style or not. We use the 4-speaker model for this experiment. Figure 4b shows naturalness preference and 4c shows percentage of the time style was transferred correctly. Our model Mix-StAGE performs best in both cases. This trend is corroborated with higher inception scores in Table 2.

Impact of Number of Speakers for Style Transfer In Table 2, we observe that increasing the number of speakers used for training also increases the average inception score for the stylized gesture generations. This is a welcome effect as it indicates increases in the diversity and the accuracy of the generations.

Attribute-level Style Transfer in Multi-Speaker Models We study four common attributes of gesture style which are also visually distinguishable by humans: (1) sitting vs. standing, (2) high vs low gesture frequency, (3) left vs right body orientation and (4) left vs right primary arm. Speakers were selected carefully to represent each extremes of these four attributes. We run a perceptual study similar to the one for speaker-level styles. However, we ask the annotators to judge if the attribute is the same in both of the videos (e.g. are both the people gesturing with the same arm?). Results from Figure 5 show that Mix-StAGE generates more (or similar) number of natural gestures with the correct attribute-level style compared to the other baselines. We also observe that it is harder for humans to determine if a person is standing or sitting, which we suspect is due to the missing waistline in the animation.



(a) Primary Arm Func. (b) Body Orientation (c) Sitting vs Standing (d) Gesture Frequency

Fig. 7: Style-Content Heatmaps for attribute-level style transfer. Each column represents the same style, while rows have input audio from different speakers. These heatmaps show that gestures are consistent across audio inputs but different between styles. Red regions correspond to the motion of the right arm, while blue corresponds to the left.

For a visual understanding of the generated gestures and stylized gestures, we plot a style-content heatmap in Figure 7, where columns represent generations for a specific style, while rows represent different speaker's audio as input. These heatmaps show that gestures are consistent across audio inputs but different between styles. Accuracy and diversity of style transfer is corroborated by inception scores in Table 2.

7 Conclusions

In this paper, we propose a new model, named Mix-StAGE, which learns a single model for multiple speakers while learning unique style embeddings for each speaker's gestures in an end-to-end manner. A novelty of Mix-StAGE was to learn a mixture of generative models conditioned on gesture style while the audio drives the co-speech gesture generation. We also introduced a new dataset, Pose-Audio-Transcript-Style (PATS), designed to study gesture generation and style transfer. It consists of 25 speakers (15 new speakers and 10 speakers from Ginosar et. al. [18]) for a total of 250+ hours of gestures and aligned audio signals. Our proposed Mix-StAGE model significantly outperformed previous state-of-the-art approach for gesture generation and provided a path towards performing gesture style transfer across multiple speakers. We also demonstrated, through human perceptual studies, that the generated animations by our model are more natural whilst being able to retain or transfer style.

Acknowledgements This material is based upon work partially supported by the National Science Foundation (Awards #1750439 #1722822), National Institutes of Health and the InMind project. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of National Science Foundation or National Institutes of Health, and no official endorsement should be inferred.

References

1. Ahuja, C., Ma, S., Morency, L.P., Sheikh, Y.: To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations. In: 2019 International Conference on Multimodal Interaction. pp. 74–84. ACM (2019) 3
2. Ahuja, C., Morency, L.P.: Language2pose: Natural language grounded pose forecasting. In: 2019 International Conference on 3D Vision (3DV). pp. 719–728. IEEE (2019) 6
3. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. pp. 3686–3693 (2014) 9
4. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. arXiv preprint arXiv:1701.07875 (2017) 5
5. Arora, S., Ge, R., Liang, Y., Ma, T., Zhang, Y.: Generalization and equilibrium in generative adversarial nets (gans). In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 224–232. JMLR. org (2017) 5
6. Bailenson, J.N., Yee, N., Merget, D., Schroeder, R.: The effect of behavioral realism and form realism of real-time avatar faces on verbal disclosure, nonverbal disclosure, emotion recognition, and copresence in dyadic interaction. Presence: Teleoperators and Virtual Environments **15**(4), 359–372 (2006) 1
7. Bansal, A., Ma, S., Ramanan, D., Sheikh, Y.: Recycle-gan: Unsupervised video retargeting. In: Proceedings of the European conference on computer vision (ECCV). pp. 119–135 (2018) 3
8. Bergmann, K., Kopp, S.: Increasing the expressiveness of virtual agents: autonomous generation of speech and gesture for spatial description tasks. In: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1. pp. 361–368 (2009) 2
9. Bian, Y., Chen, C., Kang, Y., Pan, Z.: Multi-reference tacotron by intercross training for style disentangling, transfer and control in speech synthesis. arXiv preprint arXiv:1904.02373 (2019) 3
10. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. arXiv preprint arXiv:1812.08008 (2018) 10
11. Cassell, J., Vilhjálmsson, H.H., Bickmore, T.: Beat: the behavior expression animation toolkit. In: Life-Like Characters, pp. 163–185. Springer (2004) 1
12. Chiu, C.C., Marsella, S.: Gesture generation with low-dimensional embeddings. In: Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems. pp. 781–788 (2014) 3
13. Chiu, C.C., Morency, L.P., Marsella, S.: Predicting co-verbal gestures: A deep and temporal modeling approach. In: Proceedings of the 15th international conference on Intelligent virtual agents (IVA2015). vol. 9238, pp. 152–166 (2015) 1
14. Davis, R.O., Vincent, J.: Sometimes more is better: Agent gestures, procedural knowledge and the foreign language learner. British Journal of Educational Technology **50**(6), 3252–3263 (2019) 10
15. Denton, E.L., et al.: Unsupervised learning of disentangled representations from video. In: Advances in neural information processing systems. pp. 4414–4423 (2017) 3
16. Ferstl, Y., Neff, M., McDonnell, R.: Multi-objective adversarial gesture generation. In: Motion, Interaction and Games. p. 3. ACM (2019) 1, 3
17. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015) 3
18. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3497–3506 (2019) 1, 3, 7, 8, 9, 10, 11, 14

19. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014) 7
20. Gurunath, N., Rallabandi, S.K., Black, A.: Disentangling speech and non-speech components for building robust acoustic models from found data. arXiv preprint arXiv:1909.11727 (2019) 3
21. Hao, G.Y., Yu, H.X., Zheng, W.S.: Mixgan: learning concepts from different domains for mixture generation. arXiv preprint arXiv:1807.01659 (2018) 5
22. Hasegawa, D., Kaneko, N., Shirakawa, S., Sakuta, H., Sumi, K.: Evaluation of Speech-to-Gesture Generation Using Bi-Directional LSTM Network. In: Proceedings of the 18th International Conference on Intelligent Virtual Agents (IVA18). pp. 79–86 (2018) 3
23. Hoang, Q., Nguyen, T.D., Le, T., Phung, D.: Mgan: Training generative adversarial nets with multiple generators (2018) 5
24. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 172–189 (2018) 4, 6, 8, 9, 11
25. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017) 3
26. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016) 3
27. Kendon, A.: Gesture and speech: two aspects of the process of utterance. In: M. R. Key (ed.) Nonverbal Communication and Language, pp. 207–227 (1980) 1
28. Kucherenko, T., Hasegawa, D., Henter, G.E., Kaneko, N., Kjellström, H.: Analyzing input and output representations for speech-driven gesture generation. arXiv preprint arXiv:1903.03369 (2019) 10
29. Lee, H.Y., Tseng, H.Y., Mao, Q., Huang, J.B., Lu, Y.D., Singh, M., Yang, M.H.: Drit++: Diverse image-to-image translation via disentangled representations. arXiv preprint arXiv:1905.01270 (2019) 4, 6
30. Lee, H.Y., Yang, X., Liu, M.Y., Wang, T.C., Lu, Y.D., Yang, M.H., Kautz, J.: Dancing to music. In: Advances in Neural Information Processing Systems. pp. 3581–3591 (2019) 4
31. Levine, S., Krähenbühl, P., Thrun, S., Koltun, V.: Gesture controllers. ACM Trans. Graph. **29**(4), 124:1–124:11 (Jul 2010) 3
32. Levine, S., Theobalt, C., Koltun, V.: Real-time prosody-driven synthesis of body language. ACM Trans. Graph. **28**(5), 172:1–172:10 (Dec 2009) 3
33. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in neural information processing systems. pp. 700–708 (2017) 3
34. Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks. In: Advances in neural information processing systems. pp. 469–477 (2016) 6
35. Lloyd, S.: Least squares quantization in pcm. IEEE transactions on information theory **28**(2), 129–137 (1982) 6
36. Ma, S., Mcduff, D., Song, Y.: Neural tts stylization with adversarial and collaborative games (2018) 4, 7
37. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008) 3
38. McNeill, D.: Hand and mind: What gestures reveal about thought. University of Chicago Press (1992) 1, 2, 5
39. Nagrani, A., Chung, J.S., Albanie, S., Zisserman, A.: Disentangled speech embeddings using cross-modal self-supervision. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6829–6833. IEEE (2020) 4

40. Neff, M., Kipp, M., Albrecht, I., Seidel, H.P.: Gesture modeling and animation based on a probabilistic re-creation of speaker style. ACM Transactions on Graphics (TOG) **27**(1), 1–24 (2008) 4
41. Obermeier, C., Kelly, S.D., Gunter, T.C.: A speaker's gesture style can affect language comprehension: Erp evidence from gesture-speech integration. Social cognitive and affective neuroscience **10**(9), 1236–1243 (2015) 1
42. Pelachaud, C.: Studies on gesture expressivity for a virtual agent. Speech Communication **51**(7), 630–639 (2009) 2
43. Reynolds, D.A.: Gaussian mixture models. Encyclopedia of biometrics **741** (2009) 6
44. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015) 7
45. Rosca, M., Lakshminarayanan, B., Warde-Farley, D., Mohamed, S.: Variational approaches for auto-encoding generative adversarial networks. arXiv preprint arXiv:1706.04987 (2017) 6
46. Royer, A., Bousmalis, K., Gouws, S., Bertsch, F., Mosseri, I., Cole, F., Murphy, K.: Xgan: Unsupervised image-to-image translation for many-to-many mappings. In: Domain Adaptation for Visual Understanding, pp. 33–49. Springer (2020) 6
47. Sadoughi, N., Busso, C.: Novel realizations of speech-driven head movements with generative adversarial networks. pp. 6169–6173 (04 2018). https://doi.org/10.1109/ICASSP.2018.8461967 3
48. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Advances in neural information processing systems. pp. 2234–2242 (2016) 9
49. Sargin, M.E., Yemez, Y., Erzin, E., Tekalp, A.M.: Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. IEEE Trans. Pattern Anal. Mach. Intell. **30**, 1330–1345 (2008). https://doi.org/10.1109/TPAMI.2007.70797 3
50. Shlizerman, E., Dery, L., Schoen, H., Kemelmacher, I.: Audio to body dynamics. Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (06 2018) 3
51. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1145–1153 (2017) 9
52. Smith, H.J., Cao, C., Neff, M., Wang, Y.: Efficient neural networks for real-time motion style transfer. Proceedings of the ACM on Computer Graphics and Interactive Techniques **2**(2), 1–17 (2019) 4
53. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016) 9
54. Villegas, R., Yang, J., Hong, S., Lin, X., Lee, H.: Decomposing motion and content for natural video sequence prediction. arXiv preprint arXiv:1706.08033 (2017) 3
55. Wagner, P., Malisz, Z., Kopp, S.: Gesture and speech in interaction: An overview (2014) 1
56. Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y., Saurous, R.A.: Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. arXiv preprint arXiv:1803.09017 (2018) 3, 4, 7
57. Xu, J., Gannon, P.J., Emmorey, K., Smith, J.F., Braun, A.R.: Symbolic gestures and spoken language are processed by a common neural system. Proceedings of the National Academy of Sciences **106**(49), 20664–20669 (2009) 2

58. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017) 3
59. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: Advances in neural information processing systems. pp. 465–476 (2017) 3