Zero-Shot Style Transfer for Gesture Animation driven by Text and Speech using Adversarial Disentanglement of Multimodal Style Encoding

Mireille Fares, Michele Grimaldi, Catherine Pelachaud, Nicolas Obin

Zero-Shot Style Transfer for Gesture Animation driven by Text and Speech using Adversarial Disentanglement of Multimodal Style Encoding

MIREILLE FARES, ISIR, STMS, Sorbonne University, France
MICHELE GRIMALDI, ISIR, Sorbonne University, France
CATHERINE PELACHAUD, CNRS, ISIR, Sorbonne University, France
NICOLAS OBIN, STMS, Sorbonne University, France

Modeling virtual agents with behavior style is one factor for personalizing human-agent interaction. In this paper, we propose an efficient yet effective machine learning approach to synthesize gestures driven by prosodic features and text in the style of different speakers including those unseen during training. Our model performs zero-shot multimodal style transfer driven by multimodal data from the PATS database containing videos of various speakers. We view style as being pervasive while speaking; it colors the communicative behaviors expressivity while speech content is carried by multimodal signals and text. This disentanglement scheme of content and style allows us to directly infer the style embedding even of speaker whose data are not part of the training phase, without requiring any further training or fine-tuning. The first goal of our model is to generate the gestures of a source speaker based on the *content* of two input modalities – Mel spectrogram and text semantics. The second goal is to condition the source speaker's predicted gestures on the multimodal behavior *style* embedding of a target speaker. The third goal is to allow zero-shot style transfer of speakers unseen during training without re-training the model. Our system consists of two main components: (1) a *speaker style encoder network* that learns to generate a fixed-dimensional speaker embedding *style* from a target speaker multimodal data (mel-spectrogram, pose, and text); and (2) a *sequence-to-sequence synthesis network* that synthesizes gestures based on the *content* of the input modalities - text and mel-spectrogram - of a source speaker, and conditioned on the speaker style embedding. We evaluate that our model is able to synthesize gestures of a source speaker given the two input modalities, and transfer the knowledge of target speaker style variability learned by the speaker style encoder to the gesture generation task in a zero-shot setup, indicating that the model has learned a high quality speaker representation. For our evaluation we convert the 2D generated gestures to 3D poses, and produce 3D animations of the generated gestures. We conduct objective and subjective evaluations to validate our approach and compare it with baselines.

Keywords: audio and text driven gesture synthesis, zero-shot style transfer, embodied conversational agents

1 INTRODUCTION

Human behavior style is a socially meaningful clustering of features found within and across multiple modalities, specifically in linguistic [7], spoken behavior such as the speaking style conveyed by speech prosody [29, 33], and nonverbal behavior such as hand gestures and body posture [32, 42]. Style involves the ways in which people talk differently in different situations. A same person may have different speaking styles depending on the situation (e.g. at home, at the office or with friends). These situations can carry different social meanings [5]. Different persons may also have different behavior styles while communicating in similar contexts. Style is syntagmatic. It unfolds over time in the course of an interaction and during one's life course [7]. It does not emerge unaltered from the speaker. It is continuously attuned as it is accomplished and co-produced with the audience [28]. It can be very self-conscious and at the same time can be extremely routinized to the extent that it resists attempts of being altered [28]. For instance, style-shifting has been observed in the speech of Oprah Winfrey [19], a popular host of a U.S. talk show. Internal linguistic factors such

Authors' addresses: Mireille Fares, ISIR, STMS, Sorbonne University, Paris, France, fares@isir.upmc.fr; Michele Grimaldi, ISIR, Sorbonne University, Paris, France, grimaldi@isir.upmc.fr; Catherine Pelachaud, CNRS, ISIR, Sorbonne University, Paris, France, catherine.pelachaud@sorbonne-universite.fr; Nicolas Obin, STMS, Sorbonne University, Paris, France, nobin@ircam.fr.

as lexical frequency, and external sociolinguistic factors influence the phonetic of various variables in her speech [19]. Another study [35] shows that Ellen Degeneres, another popular host of a US talk show, employs different speech styles in her TV show such as formal, consultative, casual and intimate styles. Style is specifically related to the diversity of gestures and expressivity of each specific speaker [6, 34]. All of the aforementioned points constitute a technical challenge when trying to model behavior style in virtual agents. The behavior generation model should not simply learn an overall style from multiple speakers, but should remember each speaker's specific style - idiosyncrasy - generated in a specific lexical content context and behavior expressivity. The model should be able to capture the style that are common throughout speakers, the ones that are unique to a speaker's prototypical gestures produced consciously and unconsciously, as well as the different style-shifting that may occur during speech.

Verbal and non-verbal behavior play a crucial role in sending and perceiving new information [31] in human-human interaction. Generative models that aim to predict Embodied Conversational Agents (ECA) gestures must consider the importance of producing meaningful and naturalistic gestures that are aligned with speech [9]. Non-verbal behavior must be generated and synchronized in conjunction with verbal and prosodic behavior to define their shape and time of occurrence [38]. This constitutes another technical challenge, to enable a smooth and engaging interaction between humans and ECAs by making sure that ECAs produce semantically-aware, natural, expressive and coherent gestures aligned with speech and its content.

In the present paper, we propose a novel approach to model behavior style in virtual agents and to tackle the different style modeling challenges. Our approach aims at (1) synthesizing natural and expressive upper body gestures of a source speaker, by encoding the *content* of two input modalities – text semantics and Mel spectrogram, (2) conditioning the source speaker's predicted gesture on the multimodal *style* representation of a target speaker, and therefore rendering the model able to perform style transfer across speakers, and finally (3) allowing zero-shot style transfer of newly coming speakers that were not seen by the model during training. Our model consists of two main components: first (1) a speaker style encoder network which goal is to model a specific target speaker style extracted from three input modalities – Mel spectrogram, upper-body gestures, and text semantics; and second (2) a sequence-to-sequence synthesis network that generates a sequence of upper-body gestures based on the content of two input modalities – Mel spectrogram and text semantics – of a source speaker, and conditioned on the target speaker style embedding. We trained our model on the database PATS, which was proposed in [2] and designed to study gesture generation and style transfer. It includes 3 main features that we are considering in our approach: text semantics represented by BERT embeddings, Mel spectrogram and 2D upper body poses.

1.1 Contributions

Our contributions can be listed as follows:

(1) We propose the first approach for zero-shot multimodal style transfer approach for 2D pose synthesis. At inference, an embedding style vector can be directly inferred from multimodal data (text, speech and and pose) of any speaker, by simple projection into the embedding style space (similar to the one used in [20]). The style transfer performed by our model allows the transfer of style from any unseen speakers, without further training or fine-tuning of our trained model. Thus it is not limited to the styles of the speakers of a given database. It also allows *"style preservation"* by generating gestures for multiple speakers while remembering what is unique for each speaker.

(2) To design our approach, we make the following assumptions for the separation of style and content information: *style* is possibly encoded across all modalities (text, speech, pose) and varies little or not over time; *content* is encoded only by text and speech modalities and varies over time.
(3) To implement theses assumptions, we propose an architecture for encoding and disentangling *content* and *style* information from multiple modalities. On one side, a content encoder is used to encode a content matrix from text and speech signal; on the other hand, a style encoder is used to encode a style vector from all text, speech, and signal modalities. A fader loss is introduced to effectively disentangle content and style encodings [25]. The encoding of the style takes into account 3 modalities: body poses, text semantics, and speech - Mel spectrograms. These modalities are important for gesture generation [16, 23] and are linked to style.
(4) Finally, we evaluate the 2D generated gestures by converting them to 3D poses, and simulating 3D animations of the generated gestures. The 3D poses generation is done from incomplete upper body 2D pose joints, using MocapNET, and are simulated on a 3D virtual agent. 3D poses estimation has never been done using 2D poses with such a large number of missing joints in the context of virtual agents animation.

The paper is organized as follows. The next section discusses the related work. We then describe the proposed architecture. Afterwards we describe the training regime. Then we present the objective and subjective evaluations. We finally discuss our results.

2 RELATED WORK

Since few years, a large number of gesture generative models have been proposed, principally based on sequential generative parametric models such as Hidden Markov Models HMM and gradually moving towards deep neural networks enabling spectacular advances over the last few years. Hidden Markov Models were previously used to predict head motion driven by prosody [39], and body motion [26, 27]. Chiu & Marsella [10] proposed an approach for predicting gesture labels from speech using conditional random fields (CRFs) and generating gesture motion based on these labels, using Gaussian process latent variable models (GPLVMs). These works focus on the gesture generation task driven by either one modality namely speech, or by the two modalities - speech and text. Their work focuses on producing naturalistic and coherent gestures that are aligned with speech and text, enabling a smoother interaction with ECAs, and leveraging the vocal and visual prosody. The non-verbal behavior is therefore generated in conjunction with the verbal behavior. Most of these works use a TTS for producing the voice, which, then, serves as input for computing the animation of the virtual agent. LSTM networks driven by speech were recently used to predict sequences of gestures [18] and body motions [3, 40]. LSTMs were additionally employed for synthesizing sequences of facial gestures driven by text and speech, namely the fundamental frequency (F0)[12, 13]. Generative adversarial networks (GANs) were proposed to generate realistic head motion [37] and body motions [15]. Furthermore, transformer networks and attention mechanisms were recently used for upper-facial gesture synthesis based on multimodal data - text and speech [14]. Jonell et al. [21] propose a probabilistic approach based on normalizing flows for synthesizing facial gestures in dyadic settings. Gestures driven by both acoustic and semantic information [12, 14, 24] are the closest approaches to our gesture generation task, however they cannot be used for the style transfer task.

Beyond realistic generation of human non-verbal behavior, style modelling and control in gesture is receiving more attention in order to propose more expressive behaviors that could possibly adapted to a specific audience [1, 2, 4, 11, 16, 22, 30]. Michael Neff et al.[30] propose a system that produces full body gesture animation driven by text, in the style of a specific performer. Alexanderson et al. [4] propose a generative model for synthesizing speech-driven

gesticulation, they exert directorial control over the output style such as gesture level and speed. Tero Karras et al.[22] propose a model for driving 3D facial animation from audio. Their main objective is to model the style of a single actor by using a deep neural network that outputs 3D vertex positions of meshes that correspond to a specific audio. Daniel Cudeiro et al.[11] also propose a model that synthesizes 3D facial animation driven by speech signal. Ginosar et al. [16] propose an approach for generating gestures given audio speech, however their approach uses models trained on single speakers. The aforementioned works have focused on generating nonverbal behaviors (facial expression, head movement, gestures in particular) aligned with speech [2, 11, 22, 30]. They have not consider multimodal data when modeling style, as well as when synthesizing gestures.

To our knowledge, the only attempts to model and transfer the style from multi-speakers database have been proposed by [2] and [1]. [2] presented Mix-StAGE, a speech driven approach that trains a model from multiple speakers while learning a unique style embedding for each speaker. They created PATS, a dataset designed to study various styles of gestures for a large number of speakers in diverse settings. In their proposed neural architecture, a content and a style encoder are used to extract content and style information from speech and pose. To disentangle style from content information, they assume that style is only encoded through the pose modality, and the content is shared across speech and pose modalities. A style embedding matrix whose each vector represents the style associated to a specific speaker from the training set. During training, they further propose a multimodal GAN strategy to generate poses either from the speech or pose modality. During inference, the pose is inferred by only using the speech modality and the desired style token. However, their generative model is conditioned on gesture style and driven by audio. It does not include verbal information. It cannot perform zero-shot style transfer on speakers that were not seen by their model during training. In addition, the style is associated with each unique speaker, which makes the distinction unclear between each speaker's specific style - idiosyncrasy -, the style that is shared among a set of speakers of similar settings (i.e. TV show hosts, journalists, etc...), and the style that is unique to each speaker's prototype gestures that are produced consciously and unconsciously, in addition to the different style-shifting that may occur. Moreover, the style transfer is limited to the styles of the speakers of , which prevents the transfer of style from an unseen speaker. Furthermore, the proposed architecture is based on the disentangling of content and PATS style information, which is based on the assumption that style is only encoded by gestures. However, both text and speech also convey style information, and the encoding of style must take into account all the modalities of human behavior. To tackle those issues, [1] presented a few-shot style transfer strategy based on neural domain adaptation accounting for cross-modal grounding shift between source speaker and target style. This adaptation still requires 2 minutes of the style to be transferred.

To the best of our knowledge, our approach is the first to synthesize gestures from a source speaker, which are semantically-aware, speech driven and conditioned on a multimodal representation of the style of target speakers, in a zero-shot configuration i.e., without requiring any further training or fine-tuning.

3 ZERO-SHOT MULTIMODAL STYLE TRANSFER MODEL (ZS-MSTM) FOR GESTURE ANIMATION DRIVEN BY TEXT AND SPEECH

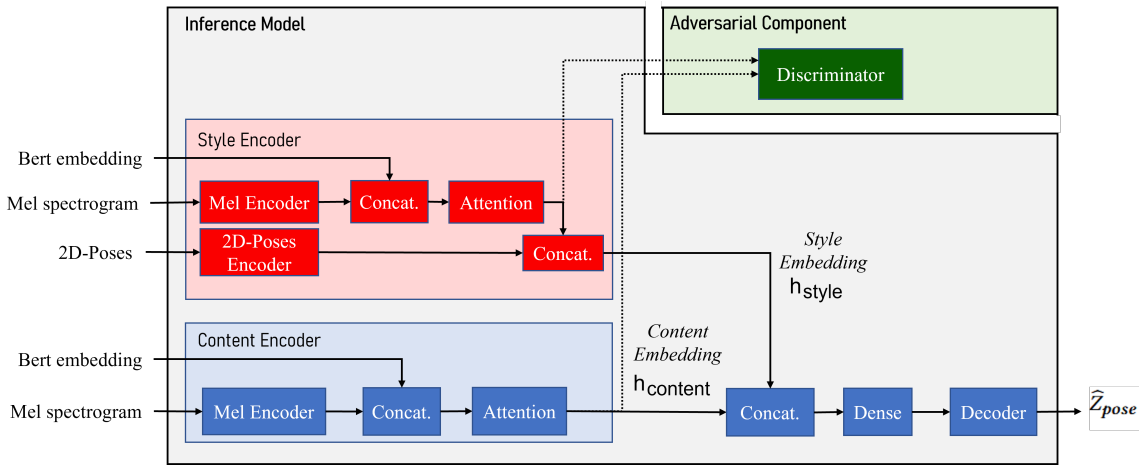


Fig. 1. **ZS-MSTM** (**Z**ero-**S**hot **M**ultimodal **S**tyle **T**ransfer **M**odel) architecture. The content encoder (further referred to as $E_{content}$) is used to encode content embedding $h_{content}$ from BERT text embeddings X_{text} and speech Mel-spectrograms Y_{speech} using a speech encoder $E^{speech}_{content}$. The style encoder (further referred to as E_{style}) is used to encode style embedding h_{style} from multimodal text X_{text}, speech Y_{speech}, and pose Z_{pose} using speech encoder E^{speech}_{style} and pose encoder E^{pose}_{style}. The generator G is a transformer network that generates the sequence of poses $\widehat{Z}_{pose}$ from the sequence of content embedding $h_{content}$ and the style embedding vector h_{style}. The adversarial module relying on the discriminator D is used to disentangle content and style embeddings $h_{content}$ and h_{style}.

We propose **ZS-MSTM** (**Z**ero-**S**hot **M**ultimodal **S**tyle **T**ransfer **M**odel), a transformer-based architecture for stylized upper-body gesture synthesis, driven by the content of a source speaker's speech - text semantics represented by BERT embeddings and audio Mel spectrogram -, and conditioned on a target speaker's multimodal style embedding. The stylized generated gestures correspond to the style of target speakers seen and unseen during training. As depicted in Fig.1, the system is composed of three main components:

(1) A **speaker style encoder** network that learns to generate a fixed-dimensional speaker embedding style from a *target speaker* multimodal data: 2D poses, BERT embeddings, and Mel spectrogram, all extracted from videos in a database.
(2) A **sequence to sequence gesture synthesis** network that synthesizes gestures based on the content of two input modalities - text embeddings and Mel spectrogram - of a *source speaker*, and conditioned on the *target speaker* style embedding. A *content encoder* is presented to encode the content of the Mel spectrogram along with BERT embeddings.
(3) An **adversarial component** in the form of a fader network [25] is used for disentangling style and content from the multimodal data.

At inference time, the adversarial component is discarded, and the model can generate different versions of poses when fed with different style embeddings. Gesturing styles for the same input speech can be altered by simply switching the style embeddings, or switching the multimodal input data fed as input to the Style Encoder.

ZS-MSTM illustrated in Fig. 1 aims at mapping multimodal speech and text feature sequences into continuous upper-body gestures, conditioned on a speaker style embedding. The network operates on the word-level: the inputs and output of the network consist of one feature vector for each word $\boldsymbol{W}$ of the input text sequence. The length of the word-level input features (text and audio) corresponds to *64* timesteps (as provided by PATS). The model generates a sequence of gestures corresponding to the same word-level features given as inputs. Gestures are sequences of 2D poses represented by X and Y positions of the joints of the skeleton. The network has an embedding dimension *d_model* equal to 768.

3.1 Content Encoder

The content encoder $E_{content}$ illustrated in Fig.1 takes as inputs BERT embedding X_{text} and audio Mel spectrograms Y_{speech} corresponding to each $\mathbf{W}$. X_{text} is represented by a vector of length 768 - BERT embedding size used in PATS. Y_{speech} is encoded using *Mel spectrogram Transformer (AST)* pre-trained *base384* model [17]. *AST* operates as follows: the input Mel spectrogram which has 128 frequency bins, is split into a sequence of 16x16 patches with overlap, and then is linearly projected into a sequence of 1D patch vectors, which is added with a positional embedding. We append a [*CLS*] token to the resulting sequence, which is then input to a *Transformer Encoder*. *AST* was originally proposed for audio classification. Since we do not intend to use it for a classification task, we remove the linear layer with sigmoid activation function at the output of the *Transformer Encoder*. We use the *Transformer Encoder*'s output of the [*CLS*] token as the Mel spectrogram representation $\mathbf{S}$. The *Transformer Encoder* has an embedding dimension equals to d_{model}, 12 encoding layers, and 12 attention heads. The word-level encoded Mel spectrogram is then concatenated with the word-level BERT embedding. A self-attention mechanism is then applied on the resulting vector. The multi-head attention layer has 4 attention heads, and an embedding size d_{att} equals to $d_{att} = d_{model} + 768$. The output of the attention layer is the vector $h_{content}$, a content representation of the source speaker's word-level Mel spectrogram and text embedding, and it can be written as follows:

$$h_{content} = sa\left(\left[E^{speech}_{content}(Y_{speech}), X_{text}\right]\right) \tag{1}$$

where: sa(.) denotes self-attention.

3.2 Style Encoder

As discussed previously, *style* is a clustering of features found within and across modalities, encompassing verbal and non-verbal behavior. It is not limited to gestural information. We consider that style is encoded in a speaker's multimodal - text, speech and pose - behavior. As illustrated in Fig.1, the style encoder E_{style} takes as input, at the word-level, Mel spectrogram Y_{speech}, BERT embedding X_{text}, and a sequence of (X, Y) joints positions that correspond to a target speaker's 2D poses Z_{pose}. *AST* is used to encode the audio input spectrogram. 3 layers of LSTMs with a hidden-size equal to d_{model} are used to encode the vector representing the 2D poses. The last hidden layer is then concatenated with the audio representation. Next, a multi-head attention mechanism is applied on the resulting vector. This attention layer has 4 attention heads and an embedding size equals to d_{att}. Finally, the output vector is concatenated with the 2D poses vector representation. The resulting vector h_{style} is the output speaker style embedding that serves to condition the network with the speaker style. The final style embedding h_{style} can therefore be written as follows :

$$h_{style} = \left[sa\left(\left[X_{text}, E^{speech}_{style}(Y_{speech})\right]\right), E^{pose}_{style}(Z_{pose})\right] \tag{2}$$

where: sa(.) denotes self-attention.

3.3 Sequence to sequence gesture synthesis

The stylized 2D poses are generated given the sequence of content representation $h_{content}$ of the source speaker's Mel spectrogram and text embeddings obtained at word-level, and conditioned by the style vector embedding h_{style} generated from a target speaker's multimodal data. For decoding the stylized 2D-poses, the sequence of $h_{content}$ and the vector h_{style} are concatenated (by repeating the h_{style} vector for each word of the sequence), and passed through a dense layer of size d_{model}. We then give the resulting vector as input to a transformer decoder. The transformer decoder is composed of $N = 1$ decoding layer, with 2 attention heads, and an embedding size equal to d_{model}. Similar to the one proposed in [41], it is composed of residual connections applied around each of the sub-layers, followed by layer normalization. Moreover, the self-attention sub-layer in the decoder stack is altered to prevent positions from attending to subsequent positions. The output predictions are offset by one position. This masking makes sure that the predictions for position index $\boldsymbol{j}$ depends only on the known outputs at positions that are less than $\boldsymbol{j}$. For the last step, we perform a permutation of the first and the second dimensions of the vector generated by the transformer decoder. The resulting vector is a sequence of 2D-poses which corresponds to:

$$\widehat{Z}_{pose} = G(h_{content}, h_{style}) \tag{3}$$

where: G is the transformer generator conditioned on latent content embedding $h_{content}$ and style embedding h_{style} . The generator loss of the transformer gesture synthesis can be written as,

$$\mathcal{L}_{rec}^{gen}(E_{content}, E_{style}, G) = \mathop{\mathbb{E}}_{\widehat{Z}_{pose}} ||\widehat{Z}_{pose} - G(h_{content}, h_{style})||_2 \tag{4}$$

3.4 Adversarial Component

Our approach of disentangling style from content relies on the fader network disentangling approach [25], where a fader loss is introduced to effectively separate content and style encodings. The fundamental feature of our disentangling scheme is to constrain the latent space of $h_{content}$ to be independent of the style embeddings h_{style}. Concretely, it means that the distribution over $h_{content}$ of the latent representations should not contain the style information. A fader network is composed of: an encoder which encodes the input information X into the latent code $h_{content}$, a decoder which decodes the original data from the latent, and an additional variable h_{style} used to condition the decoder with the desired information (a face attribute in the original paper). The objective of the fader network is to learn a latent encoding $h_{content}$ of the input data that is independent on the conditioning variable h_{style} while both variables are complementary to reconstruct the original input data from the latent variable $h_{content}$ and the conditioning variable h_{style}. To do so, a discriminator D is optimized to predict the variable h_{style} from the latent code $h_{content}$; on the other side the auto-encoder is optimized using an additional adversarial loss so that the classifier D is unable to predict the variable h_{style}. Contrary to the original fader network in which the conditional variable is discrete within a finite binary set (0 or 1 for the presence or absence attribute), in this paper the conditional variable h_{style} is continuous. We then formulate this discriminator as a regression on the conditional variable h_{style}: the discriminator learns to predict the style embedding h_{style} from the content embedding $h_{content}$, as:

$$\widehat{h}_{style} = D(h_{content}) \tag{5}$$

While optimizing the discriminator, the discriminator loss $\mathcal{L}^{dis}$ must be as low as possible, such as:

$$\mathcal{L}^{dis}(D) = \mathop{\mathbb{E}}_{\widehat{h}_{style}} ||h_{style} - D(h_{content})||_2 \tag{6}$$

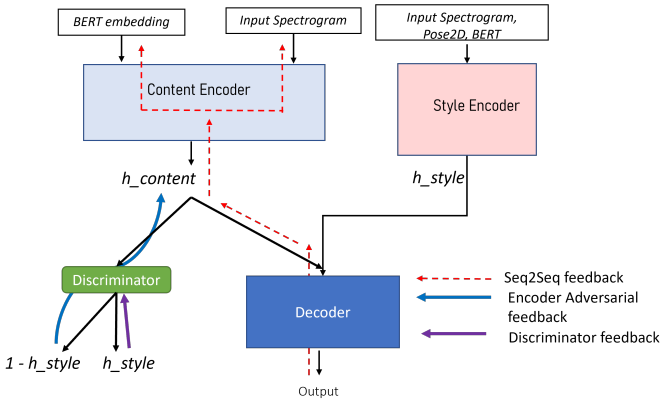


Fig. 2. Fader network for multimodal content and style disentangling.

In turn, optimizing the generator loss including the fader loss $\mathcal{L}_{adv}^{gen}$, the discriminator must not be able to predict correctly the style embedding h_{style} from the content embedding $h_{content}$ conducting to a high discriminator error and thus a low fader loss. The adversarial loss can be written as,

$$\mathcal{L}_{adv}^{gen}(E_{content}, E_{style}, G) \underset{\widehat{h}_{style}}{\mathbb{E}} ||1 - (h_{style} - D(h_{content}))||_2 \tag{7}$$

To be consistent, the style prediction error is preliminary normalized within 0 and 1 range.
Finally, the total generator loss can therefore be written as follows:

$$\mathcal{L}_{total}^{gen}(E_{content}, E_{style}, G) = \mathcal{L}_{rec}^{gen}(E_{content}, E_{style}, G) + \lambda \mathcal{L}_{adv}^{gen}(E_{content}, E_{style}, G) \tag{8}$$

where λ is the adversarial weight that starts off at 0 and is linearly incremented by 0.01 after each training step. The discriminator D and the generator G are then optimized alternatively as described in [25].

4 TRAINING

This section describes the training regime we follow for our model. We trained our network using the PATS dataset [2]. PATS was created to study various styles of gestures. The dataset contains upper-body 2D pose sequences aligned with corresponding Mel spectrogram, and BERT embeddings. It offers 251 hours of data, with a mean of 10.7 seconds and a standard deviation of 13.5 seconds per interval. PATS gathers data from 25 speakers with different behavior styles from various settings (e.g., lecturers, TV shows hosts). It contains also several annotations. The spoken text has been transcribed in PATS and aligned with the speech. The 2D body poses have been extracted with OpenPose. Each speaker is represented by their lexical diversity and the spatial extend of their arms. While in PATS arms and fingers have been extracted, we do not consider finger data in our work. That is we do not model and predict 2D finger joints. This choice arises as the analysis of finger data is very noisy and not very accurate.

We consider two test conditions: *Seen Speaker* and *Unseen Speaker*. The *Seen Speaker* condition aims to assess the style transfer correctness that our model can achieve when presented with speakers that were seen during training as target style. On the other hand, the *Unseen Speaker* condition aims to assess the performance of our model when presented with unseen target speakers, to perform zero-shot style transfer. Seen and unseen speakers are specifically selected from PATS to cover a diversity of stylistic behavior with respect to lexical diversity and spatial extent as

reported by [2][1]. For each PATS speaker, there is a train, validation and test set already defined in the database. For testing the *Seen Speaker* condition, our training set includes the train sets of 16 PATS speakers: "Shelly", "Jon", "Fallon", "Bee", "Ellen", "Oliver", "Lec_cosmic", "Lec_hist", "Seth", "Conan", "Angelica", "Rock", "Noah", "Ytch_prof", "Lec_law", and "Ytch_dating". Six other speakers are selected for the *Unseen Speaker* condition, and their test sets are also used for our experiments. These six speakers "Lec_evol", "Almaram", "Huckabee", "Ytch_charisma", "Minhaj", and "Chemistry" differ in their behavior style and lexical diversity. Each training batch contains 24 pairs of word embeddings, Mel spectrogram, and their corresponding sequence of (X, Y) joints of the skeleton (of the upper-body pose). We use Adam optimizer with $\beta_1 = 0.95, \beta_2 = 0.999$. For balanced learning, we use a scheduler with an initial learning rate of 0.00001, with *warmup steps = 20000*. We train the network for 200 epochs. All features values are normalized so that the dataset mean and standard deviation are 0 and 0.5, respectively.

5 3D POSE GENERATION AND SIMULATION

Previous evaluation studies of models learned from video data have used 2D stick figures for their subjective evaluation [2]. Even when the 2D stick figure is projected on the video of a human speaker, the animation is not always readable as, in particular, it is missing information on the body pose in the Z direction (the depth axis). So we choose to convert the 2D poses into 3D poses. We visualize the behavior animation resulting from our model on a 3D virtual agent. As in [2], we train our model on the database PATS, and therefore the generated 2D body poses correspond to incomplete skeleton joints; missing joints include lower body joints, as well as torso joints. To visualize the resulting animations of our model, we convert the 2D poses into 3D poses and use 3D human mesh.

We develop an approach that generates 3D poses from incomplete upper body 2D pose joints using MocapNET [36], an ensemble of SNN encoders that estimates the 3D human body pose based on 2D joint estimations extracted from monocular RGB images. It outputs skeletal information directly into the BVH format which can be rendered in real-time or imported without any additional processing in most popular 3D animation software. MocapNET operates on 2D joint input, received in the popular COCO[8] or BODY25[8] format. In order to be used, the file containing the predictions are formatted following the BODY25 format and the 2D joints are mapped to respect the BODY25 joints. The JSON files with 2D detections are subsequently converted to CSV files and then to 3D BVH files using the MocapNET. Finally we add zeros for the missing joints. MocapNET is trained using a 1920x1080 "virtual camera" to emulate a GoPRO Hero 4 running at the Full-HD mode. We adapted the output of our gesture generation model to such a configuration. We also set up the frames resolution to correspond to the original video stream size. Once the BVH file is created we use the 3D animation software Blender to simulate the animation. Finally, we apply a 3D human mesh to the skeleton to simulate a 3D human animation. The mesh is taken from Mixamo[2], an online database of characters and mocap animations used in art projects, movies and games. In order to fuse the mesh with the skeleton, we scale the mesh to fit the skeleton and we parent the skeleton and the mesh with automatic weights.

6 EVALUATION METRICS AND STUDIES

To measure the performance of our work, we conducted several objective and subjective evaluation studies we present in this section. We start by introducing the metrics we use for the objective studies; we follow by explaining the protocol we follow for the perceptive studies as well as the creation of stimuli we use.

[1]https://chahuja.com/pats/
[2]https://www.mixamo.com/

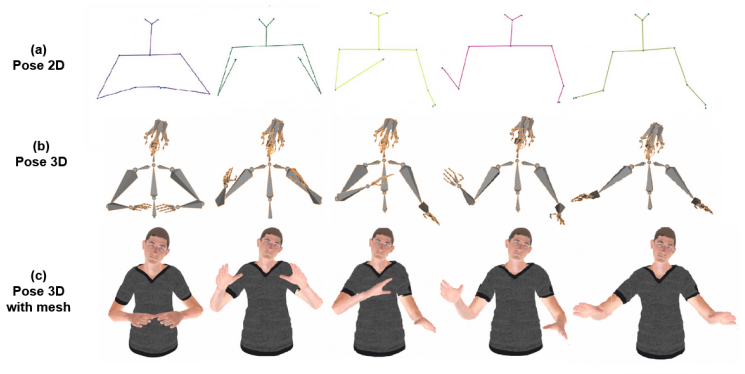


Fig. 3. A sequence of gestures corresponding to a sequence of 2D poses. (a) 2D poses. (b) The corresponding sequence of 3D poses computed by MocapNet and simulated with Blender. (c) Resulting animation with a 3D human mesh.

6.1 Objective Evaluation Metrics

In our work, we have defined style by the behavior expressivity of a speaker. To evaluate objectively our works, we define metrics to compare the behavior expressivity of a speaker from the ground truth with the predicted behavior expressivity generated by our model in different conditions. Following works on behavior expressivity by [34, 43], we define 4 objective behavior dynamics metrics to evaluate the style transfer of different target speakers: Acceleration, jerk and velocity that are averaged over the values of all upper-body joints, as well as the speaker's average bounding box perimeter of his/her body movements extension. In addition, we compute the acceleration, jerk and velocity of only the left and right wrists, to obtain information on the arms movements expressivity [23, 43].

6.2 Subjective Evaluation

We also conduct three human perceptual studies. As a pre-evaluation of our approach, we conduct a human perceptual study (**Study 1**) to validate the 2D to 3D pose conversion by measuring the *resemblance* of the 3D animations to the ground truth in terms of the expressivity of the style (gesture amplitude and dynamics), in addition to the quality of the produced 3D animations (such as naturalness and comprehensibility of movements). Then, to investigate human perception of the stylized upper-body gestures produced by our model, we conduct another human perceptual study (**Study 2**) that aims to evaluate the gestures produced by our model and its capacity to perform "style preservation". A third study (**Study 3**) was conducted to evaluate the style transfer of speakers seen during training - *Seen Speaker* condition - , as well as speakers unseen during training - *Unseen Speaker* condition. In these studies, we present a virtual agent simulated with the converted 3D poses of the 2D poses synthesized by our model. **Study 3** aims to assess the *resemblance* of the produced stylized gestures to the target style. We additionally compare in **Study 3** our model's produced stylized gestures in *Seen Speaker* condition, to Mix-StAGE that we consider our baseline. We used 7 factors linked to behavior expressivity to assess the quality of the 3D animation. We follow the recommendations proposed in [44] and assess on a 1 to 7 likert scale the first 5 factors: *naturalness*, *coherence*, *human-likeness*, *appropriateness*, and *comprehensibility*. We add the 2 other factors *synchronization*, and *alignment* to evaluate the gestures' temporal property with speech. In addition, 3 factors are used to evaluate the *resemblance*, *resemblance in terms of gestures amplitude*, and *resemblance in terms of gestures dynamics* between the human gestures and the virtual agent's gestures. We note that we distinguish between two types of factors that we want to assess in our studies: the first ones (7 expressivity

factors) are related to evaluating the virtual agent's *behavioral expressivity* (for **Study 1** and **Study 2**), and the second ones (3 resemblance factors) are to assess the *resemblance* of our model's stylized produced gestures with the ground truth (for **Study 1**) and with the target style (for **Study 3**). Each factor is rated on a 7 likert scale. 30 participants are recruited for each study, including for the pre-evaluation study (**Study 1**), on Prolific, an online crowd-sourcing website. Participants are selected such that they are fluent in English and have a university degree. Attention checks are added in the beginning and the middle of each study to filter out inattentive participants. All the animations presented in these studies are produced on a 3D virtual agent.

6.2.1 ***Study 1: 3D Animation Pre-Evaluation.*** The first human perceptual study we conduct aims to assess our approach for the 2D to 3D pose conversion. In this study, we present 4 pairs of videos: for each pair, the first video shows the generated 3D poses simulated on a virtual agent, and the second one is the video of the original speaker performing the same gestures. The 2D poses that we use for this 3D conversion are ground truth data extracted from PATS.

6.2.2 ***Study 2: Gesture Generation Evaluation.*** To assess the quality of the 2D poses generated by our model, and its ability to perform "style preservation" and remember the unique style of each speaker, we conduct another human perceptual study. We use the 7 expressivity factors that are used in the pre-evaluation study to assess the quality of the produced virtual agent's gestures. This study consists of 8 videos: 4 videos show 3D animations of our model's predictions, and 4 other videos show the converted 2D to 3D poses animation of the original speaker's gestures which serve as ground truth. For each video, participants are asked to rate the 7 expressivity factors on a 1 to 7 likert scale [44].

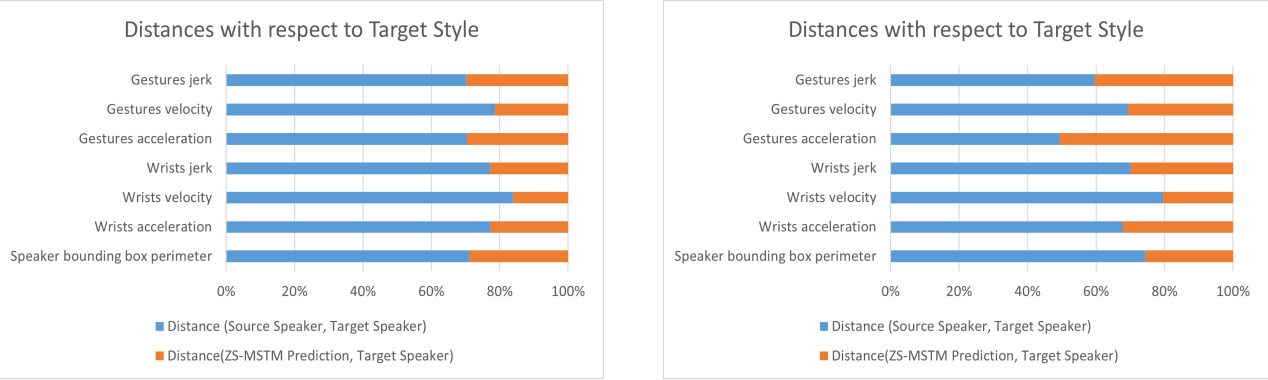
6.2.3 ***Study 3: Style Transfer and Zero-Shot Style Transfer Evaluation.*** The third perceptive study aims to assess the style transfer correctness performed by our model for both conditions: *Seen Speaker* and *Unseen Speaker*. For each condition, participants watch 3 videos representing the ground truth (*video 1*), the target speaker (*video 2*) and our model (*video 3*), respectively. We ask the participants to answer questions related to the 3 resemblance factors, provided in a random order. For the *Seen Speaker* condition, we present 12 videos: 4 videos show the 3D animation of the source speaker gestures, 4 other videos show the 3D animation of the target speaker gestures, and the remaining 4 videos show the simulation of our model's predictions in 3D, after performing the style transfer from the target speaker to the source speaker. For the *Unseen Speaker* condition, we present 9 videos (different videos from the above ones): 3 videos with the source speaker gestures, 3 with the target speaker gestures, and the remaining 3 with our model's 3D simulated predictions after performing the style transfer from target speakers not seen during training, to the source speakers. We note that in this experimental study, the *resemblance* factors are the most important ones, since we want to assess the degree of resemblance of our model's stylized gestures to the target style. For each set of questions in each condition, the target 3D animation is presented to the participants as a "baseline". We ask the participants to choose one of the two video - the source speaker 3D animation, and the 3D simulation of our model's predictions - that resembles the most to the baseline in terms of the 3 resemblance factors. Participants are asked the following questions: (1) Which video resembles the most to the baseline video ?; (2) Which video resembles the most to the baseline video in terms of gestures dynamics ?; and (3) Which video resembles the most to the baseline video in terms of gestures amplitude ?

Comparing to the baseline **Mix-StAGE:** We compare our stylized generated gestures in *Seen Speaker* condition with the predictions of *Mix-StAGE*[2], which serves as a baseline for this condition. We ask the participants to watch 3 videos representing the ground truth (*video 1*), the target speaker (*video 2*), and Mix-StAGE predictions after performing

style transfer from target speakers to source speakers (*video 3*). We repeat this question 3 times (presenting 9 videos in total), and assess the *resemblance* of *Mix-StAGE*'s produced gestures with respect to the target speakers.

7 RESULTS AND DISCUSSION

7.1 Objective Evaluation Results



(a) style transfer from seen speakers

(b) style transfer from unseen speakers

Fig. 4. Distances between the target speaker style and each of the source style and ZS-MSTM's generated gestures style for seen target speakers: on left, (a) style transfer using seen speaker during training; on right, (b) style transfer using unseen speaker during training.

Objective evaluation experiments are conducted for evaluating the performance of our model in the *Seen Speaker* and *Unseen Speaker* conditions. For *Seen Speaker* condition, experiments are conducted on the test set that includes the 16 speakers that are seen by our model during training. For *Unseen Speaker* condition, experiments are also conducted on another test set that includes the 6 speakers that were not seen during training. For both conditions, we define two sets of distances: ***Dist.(****Source*, *Target***)** - representing the average distance between the source style and the target style on the corresponding test set -, and ***Dist.(****ZS-MSTM*, *Target***)** - representing the average distance between our model's gestures style and the target style -. Fig. 4 (a) reports the experimental results on the *Seen Speaker* test set. It illustrates the results of ***Dist.(****Source*, *Target***)** in terms of behaviors dynamics and speaker bounding box perimeter between the target speaker style and the source speaker style. Experimental results for the *Unseen Speaker* test set are depicted in Fig. 4 (b). For *Seen Speaker* condition (Fig. 4(a)), ***Dist.(****Source*, *Target***)** is higher than 70% of the total distance for all behavior dynamics metrics.; thus ***Dist.(****ZS-MSTM*, *Target***)** is less than 30% of the total distance for all behavior dynamics metrics. Wrists velocity, jerk and acceleration results reveal that the virtual agent's arms movements show the same expressivity dynamics as the target style (***Dist.(****ZS-MSTM*, *Target***)** < 22%).The style transfer from target speaker "Shelly" to source speaker "Angelica" - knowing that Angelica is a *Seen Speaker* - shows that the distance of predicted gestures' behavior dynamics metrics are close (distance < 20%) to "Shelly" (*target style*), while the ones between "Angelica" and "Shelly" are far (distance > 80%). The perimeter of the prediction's bounding box (BB) is closer (distance < 30 %) to the target speaker's BB perimeter than the source . The closeness between predictions dynamics behavior metrics values are shown for all speakers in the *Seen Speaker* condition, specifically for the following style transfers - *target* to *source* -

: "Fallon" to "Shelly", "Bee" to "Shelly", "Conan" to "Angelica", "Oliver" to "lec_cosmic", which are considered having different lexical diversity, as well as spatial average extent, as reported by the authors of PATS [2].

For the *Unseen Speaker* condition, results reveal that our model is capable of reproducing the style of the 6 unseen speakers. As depicted in Fig. 4 (b), for all behavior dynamics metrics, as well as the bounding box perimeter, ***Dist.(****Source*, *Target***)** is higher than 50% of the total distances for all metrics. Results show that for wrists velocity, jerk and acceleration, ***Dist.(****ZS-MSTM*, *Target***)** is less than 33%. Thus, arm movement's expressivity produced by *ZS-MSTM* is close to the one of the target speaker style. Moreover, the perimeter of the prediction's bounding box is close (distance < 30 %) to the target speaker's, while the distance between the BB perimeter of the source and the target is far (distance > 70 %). While our model has not seen "Lec_evol"'s multimodal data during training, it is yet capable of transferring his behavior expressivity style to the source speaker "Oliver". It is also capable of performing zero-shot style transfer from the target speaker "Minhaj" to the source speaker "Conan". In fact, results show that wrists acceleration and jerk values of our model's generated gestures are very close to those of the target speaker "Minhaj". We observe the same results for the 6 speakers for the *Unseen Speaker* condition.

We additionally conduct a Fisher's LSD Test to do pair-wise comparisons on all metrics, for the two set of distances - ***Dist.(****Source*, *Target***)**, and ***Dist.(****ZS-MSTM*, *Target***)** - in both conditions. We find significant results ($p < 0.003$) for all distances in both conditions.

7.2 Human Perceptual Studies Results

7.2.1 ***Study 1: 3D Animation Pre-Evaluation.*** The first human perceptual study is the pre-evaluation to assess the 3D data animation and simulation on a virtual agent, which are converted from the 2D generated poses. We calculate the mean values obtained on the 7 expressivity factors and on the 3 resemblance factors. Results show that all factors received a mean score above 3 on a likert-scale from 1 to 7. They reveal that the 2D to 3D conversion of the 2D-poses generated by our model tend to resemble the human's gestures which served as ground truth in this evaluation. We observe that the factor *Resemblance* gets the highest mean (above 4) and that the factor *Gestures Amplitude Resemblance* gets the highest second mean score, followed by the factor *Naturalness*. This indicates that the 3D animations show gestures that resemble the human's gestures, especially in terms of gestural amplitude resemblance. We obtain similar mean scores (3.5< *mean* <3.6) for the factors *Comprehensibility*, *Gestures Dynamics Resemblance*, *Likeness*, and *Alignment*. The mean score for the remaining factors is 3.1. While the 3D pose animation has not received the highest possible rate, its results are nevertheless good enough to be used as ground truth. In the remaining evaluations, all animations are obtained with this method, offering similar behavior quality.

7.2.2 ***Study 2: Gesture Generation Evaluation.*** The second human perceptual study consists of assessing the quality of the generated poses and the ability of our model to perform "style preservation", thus its capacity of remembering the unique style of each speaker. We calculate the mean scores for the 7 behavioral expressivity factors. We observe that our model's predictions (**P**) get mean values that are close to those of the ground truth (**GT**), especially for the factors *Appropriateness* (mean difference(**GT**, **P**)=0.1) and *Comprehensibility* (mean difference(**GT**, **P**)=0.3). The remaining factors have higher mean difference between the ground truth and predictions: *Coherence* (mean difference=0.4), *Human-likeness* (mean difference=0.44), *Synchronization* (mean difference=0.5), *Alignment* (mean difference=0.51), and *Synchronization* (mean difference=0.53). We additionally perform a Fisher's LSD Test to do pair-wise comparisons of the means of the 7 factors. Significant results ($p < 0.001$) are found for the factors *Appropriateness*, *Comprehensibility*, *Coherence* and *Human-Likeness* when comparing values for the Ground Truth gestures those of our model's generated

Resemblance Metrics	ZS-MSTM - Seen Speaker		ZS-MSTM - Unseen Speaker		Mix-StAGE	
Resemblance to the target style	***Source Style***	***Prediction***	***Source Style***	***Prediction***	***Source Style***	***Prediction***
Globally	0.35 ± 0.02	0.65 ± 0.04	0.46 ± 0.01	0.54 ± 0.03	0.57 ± 0.03	0.43 ± 0.04
W.r.t. gesture dynamics	0.32 ± 0.05	0.68 ± 0.05	0.47 ± 0.02	0.53 ± 0.05	0.56 ± 0.03	0.44 ± 0.04
W.r.t. gesture amplitude	0.42 ± 0.03	0.58 ± 0.06	0.42 ± 0.04	0.58 ± 0.04	0.54 ± 0.05	0.46 ± 0.05

Table 1. Results of the perceptual study for the conditions ZS-MSTM (seen speakers), ZS-MSTM (unseen speakers), and baseline (Mix-StAGE). We also report the confidence intervals.

gestures. This constitutes experimental validation that our model is perceived significantly close to the ground truth, and therefore allows "style preservation". Therefore, our model is able to remember the unique style of each speaker, even though it is trained on multiple ones. While our model is perceived significantly close to the ground truth, results show that we still need to leverage the synchronization of the produced gestures with the speech and its content.

7.2.3 ***Study 3: Style Transfer and Zero-Shot Style Transfer Evaluation***. The first four columns (**ZS-MSTM - Seen Speaker** and **ZS-MSTM - Unseen Speaker**) of Table 1 shows the results of the human perceptual study for assessing the stylized gestures generated by our model for both conditions *Seen Speaker* and *Unseen Speaker*. Results show that, on a scale from 0 to 1 representing the number of times our model is selected to resemble the target style, our model's predictions get values above 0.58 for condition *Seen Speaker*, and values between 0.53 and 0.58 for condition *Unseen Speaker*. Our model's generated style in condition *Unseen Speaker* is perceived as having quite high resemblance to the target style (score of 0.54), especially in terms of gesture amplitude (score of 0.53) and gesture dynamics (score of 0.58). We additionally performed t-test comparison between source style values and prediction style scores for the conditions *Unseen Speaker* and *Seen Speaker*. Significant results ($p < 0.001$) are found between the Source scores and the Prediction scores. These results reveal that our model's generated stylized gestures are significantly perceived as being closer to the target style than to the source style.

Comparing to the baseline **Mix-StAGE:** The first two columns (**ZS-MSTM - Seen Speaker**) and the last two columns (**Mix-StAGE**) of Table 1 present the results when comparing our generated gestures in condition *Seen Speaker* with the baseline *Mix-StAGE* which only operates in this condition and not in the condition *Unseen Speaker*. On a scale from 0 to 1 representing the number of times our model is selected to resemble the target style, our model gets scores between 0.58 and 0.65, while Mix-StAGE gets lower scores, between 0.43 and 0.46. We additionally conduct a Fisher LSD test to do pair-wise comparisons of the means between the 3 factors of both conditions *Mix-StAGE* and *ZT-MSTM*, and identify the cases where the means are statistically different. We find a significant difference ($p < 0.003$) for the factor *Resemblance in terms of gesture dynamics*, and *Resemblance in terms of gesture amplitude.*

8 CONCLUSION AND FUTURE WORK

We have presented the first approach for zero-shot multimodal style transfer for 2D pose synthesis that allows the transfer of style from any speakers unseen during the training phase. To visualize the resulting animation of a virtual agent, we have developed a model that converts 2D poses into 3D poses. Adding human mesh on the 3D poses allows us to simulate the 3D behavior of a virtual agent. Objective and subjective evaluations show that our model produces stylized animations that are close to the target speakers style even for unseen speakers. We have evaluated our model using behavior expressivity metrics as well as perceptive factors. In a next future, we plan to expand our model to consider dialog acts, and other semantic information to model more specifically gesture types (deictic, iconic, and metaphoric). In addition, we want to extend our style model to cover facial expressions and head movements.

REFERENCES

[1] Chaitanya Ahuja, Dong Won Lee, and Louis-Philippe Morency. 2022. Low-Resource Adaptation for Personalized Co-Speech Gesture Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[2] Chaitanya Ahuja, Dong Won Lee, Yukiko I Nakano, and Louis-Philippe Morency. 2020. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *European Conference on Computer Vision*. Springer, 248–265.

[3] Chaitanya Ahuja, Shugao Ma, Louis-Philippe Morency, and Yaser Sheikh. 2019. 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*. 74–84.

[4] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 487–496.

[5] Allan Bell. 1984. Language style as audience design. *Language in society* 13, 2 (1984), 145–204.

[6] Kirsten Bergmann and Stefan Kopp. 2009. 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*. 361–368.

[7] Kathryn Campbell-Kibler, Penelope Eckert, Norma Mendoza-Denton, and Emma Moore. 2006. The elements of style. In *Poster presented at New Ways of Analyzing Variation*, Vol. 35.

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.

[9] J. Cassell. 2000. Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. In *Embodied Conversational Characters*, S. Prevost J. Cassell, J. Sullivan and E. Churchill (Eds.). MITpress, Cambridge, MA.

[10] Chung-Cheng Chiu and Stacy Marsella. 2014. Gesture generation with low-dimensional embeddings. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 781–788.

[11] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. 2019. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10101–10111.

[12] Mireille Fares. 2020. Towards multimodal human-like characteristics and expressive visual prosody in virtual agents. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 743–747.

[13] Mireille Fares, Catherine Pelachaud, and Nicolas Obin. 2021. Multimodal-Based Upper Facial Gestures Synthesis for Engaging Virtual Agents. In *WACAI 2021*.

[14] Mireille Fares, Catherine Pelachaud, and Nicolas Obin. 2021. Multimodal generation of upper-facial and head gestures with a Transformer Network using speech and text. *arXiv preprint arXiv:2110.04527* (2021).

[15] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2019. Multi-objective adversarial gesture generation. In *Motion, Interaction and Games*. 1–10.

[16] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning Individual Styles of Conversational Gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[17] Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778* (2021).

[18] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 79–86.

[19] Jennifer Hay, Stefanie Jannedy, and Norma Mendoza-Denton. 1999. Oprah and/ay: Lexical frequency, referee design and style. In *Proceedings of the 14th international congress of phonetic sciences*. University of California Berkeley, CA, 1389–1392.

[20] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems* 31 (2018).

[21] Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. 2020. Let's Face It: Probabilistic Multi-modal Interlocutor-aware Generation of Facial Gestures in Dyadic Settings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[22] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–12.

[23] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 97–104.

[24] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the ACM International Conference on Multimodal Interaction*.

[25] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Fader networks: Manipulating images by sliding attributes. *Advances in neural information processing systems* 30 (2017).

[26] Sergey Levine, Christian Theobalt, and Vladlen Koltun. 2009. Real-time prosody-driven synthesis of body language. In *ACM SIGGRAPH Asia 2009 papers*. 1–10.

[27] Stacy Marsella, Ari Shapiro, Andrew Feng, Yuyu Xu, Margaux Lhommet, and Stefan Scherer. 2013. Towards higher quality character performance in previz. In *Proceedings of the Symposium on Digital Production*. 31–35.

[28] Norma Mendoza-Denton. 1999. Style. *Journal of Linguistic Anthropology* 9, 1/2 (1999), 238–240.

[29] Sungwoo Moon, Sunghyun Kim, and Yong-Hoon Choi. 2022. MIST-Tacotron: End-to-End Emotional Speech Synthesis Using Mel-Spectrogram Image Style Transfer. *IEEE Access* 10 (2022), 25455–25463.

[30] Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics (TOG)* 27, 1 (2008), 1–24.

[31] Sigrid Norris. 2004. *Analyzing multimodal interaction: A methodological framework.* Routledge.

[32] Christian Obermeier, Spencer D Kelly, and Thomas C Gunter. 2015. A speaker's gesture style can affect language comprehension: ERP evidence from gesture-speech integration. *Social cognitive and affective neuroscience* 10, 9 (2015), 1236–1243.

[33] Nicolas Obin. 2011. *MeLos: Analysis and modelling of speech prosody and speaking style.* Ph.D. Dissertation. Université Pierre et Marie Curie-Paris VI.

[34] Catherine Pelachaud. 2009. Studies on gesture expressivity for a virtual agent. *Speech Communication* 51, 7 (2009), 630–639.

[35] Eric Trio Putra and Rusdi Noor Rosa. 2019. The analysis of speech style used by Ellen Degeneres in Ellen talk show. *English Language and Literature* 8, 3 (2019).

[36] Ammar Qammaz and Antonis A Argyros. 2019. MocapNET: Ensemble of SNN Encoders for 3D Human Pose Estimation in RGB Images. In *British Machine Vision Conference (BMVC 2019)*. BMVA, Cardiff, UK. http://users.ics.forth.gr/argyros/res_mocapnet.html

[37] Najmeh Sadoughi and Carlos Busso. 2018. Novel realizations of speech-driven head movements with generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6169–6173.

[38] M Salem, K Rohlfing, S Kopp, and F Joublin. 2011. A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. In *RoMan*. IEEE, 247–252.

[39] Mehmet E Sargin, Yucel Yemez, Engin Erzin, and Ahmet M Tekalp. 2008. Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 8 (2008), 1330–1345.

[40] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. 2018. Audio to body dynamics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7574–7583.

[41] A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A N Gomez, L Kaiser, and I Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[42] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. , 209–232 pages.

[43] H.G. Wallbott. 1998. Bodily expression of Emotion. *European Journal of Social Psychology* 28 (1998), 879–896.

[44] Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. 2022. A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems* (2022).