Zero-shot style transfer for gesture animation driven by text and speech using adversarial disentanglement of multimodal style encoding

Mireille Fares, Catherine Pelachaud, Nicolas Obin

frontiers

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

Mireille Fares [1,*], **Catherine Pelachaud** [3] **and Nicolas Obin** [4]

[1] *ISIR, STMS, Sorbonne University, Paris, France*
[2] *ISIR, Sorbonne University, Paris, France,*
[3] *CNRS, ISIR, Sorbonne University, Paris, France,*
[4] *STMS, Sorbonne University, Paris, France*

Correspondence*:
Mireille Fares
fares@isir.upmc.fr, fares@ircam.fr

ABSTRACT

Modeling virtual agents with behavior style is one factor for personalizing human-agent interaction. 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. We conduct objective and subjective evaluations to validate our approach and compare it with baselines.

Keywords: multimodal gesture synthesis, zero-shot style transfer, embodied conversational agents, multimodal behavior style, transformers

1 INTRODUCTION

Embodied Conversational Agents are virtually embodied agents that are capable of autonomously communicating with people in a socially intelligent manner using multimodal behaviors (Lugrin (2021)). The field of research in *ECAs* has emerged as new interface between humans and machines. ECAs behaviors are often modeled from human communicative behaviors. They are endowed with the capacities to recognize and generate verbal and non-verbal cues (Lugrin (2021)), and are envisioned to support humans in their daily lives. This work revolves around modeling multimodal data and learning the complex correlations between the different modalities employed in human communication. More specifically, the objective is to model the multimodal ECAs' behavior with their *behavior style*.

Human *behavior style* is a socially meaningful clustering of features found within and across multiple modalities, specifically in *linguistic* (Campbell-Kibler et al. (2006)), *spoken behavior* such as the speaking style conveyed by speech prosody (Moon et al. (2022); Obin (2011)), and *nonverbal behavior* such as hand gestures and body posture (Obermeier et al. (2015); Wagner et al. (2014)).

Behavior 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 (Bell (1984)). Different persons may also have different behavior styles while communicating in similar contexts. *Behavior style* is syntagmatic. It unfolds over time in the course of an interaction and during one's life course (Campbell-Kibler et al. (2006)). It does not emerge unaltered from the speaker. It is continuously attuned, accomplished and co-produced with the audience (Mendoza-Denton (1999)). 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 (Mendoza-Denton (1999)). Movements and gestures are person-specific and *idiosyncratic* in nature (McNeill et al. (2005)), and each speaker has his or her own non-verbal behavior style that is linked to his/her personality, role, culture, etc.

A large number of generative models were proposed in the past few years for synthesizing gestures of ECAs. Style modelling and control in gesture is receiving attention in order to propose more expressive ECAs behaviors that could possibly be adapted to a specific audience (Neff et al. (2008); Karras et al. (2017); Cudeiro et al. (2019); Ahuja et al. (2020); Ginosar et al. (2019a); Alexanderson et al. (2020); Ahuja et al. (2022)). They assume that *behavior style* is encoded in the *body gesturing*. Some of these works generate full body gesture animation driven by text in the style of one specific speaker (Neff et al. (2008)). Other approaches (Alexanderson et al. (2020); Karras et al. (2017); Cudeiro et al. (2019); Ginosar et al. (2019a)) are speech-driven. For some of these approaches, the behavior style of the synthesized gestures is changed by exerting direct control over the synthesized gestures' velocity and force (Alexanderson et al. (2020)). For others (Cudeiro et al. (2019); Karras et al. (2017); Ginosar et al. (2019a)), they produce the gestures in the style of a *single speaker* by training their generative models on one *single speaker*'s data, and synthesizing the gestures corresponding to this specific speaker's audio. Moreover, verbal and non-verbal behavior plays a crucial role in communication in human-human interaction (Norris (2004)). Generative models that aim to predict communicative gestures of ECAs must produce expressive semantically-aware gestures that are aligned with speech (Cassell (2000)).

We propose a novel approach to model *behavior style* in ECAs and to tackle the different *behavior style* modeling challenges. We view *behavior style* as being pervasive while speaking; it colors the communicative behaviors expressivity while speech content is carried by multimodal signals and text. 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. 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. The disentanglement scheme of *content* and *style* allows us to directly infer the style embedding even of speakers whose data are not part of the training phase, without requiring any further training or fine-tuning.

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. Our model is trained on the *multi-speaker* database PATS, which was proposed in Ahuja et al. (2020) and designed to study gesture generation and style transfer. It includes 3 main modalities that we are considering in our approach: text semantics represented by BERT embeddings, Mel spectrogram and 2D upper body poses.

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 poses) of any speaker, by simple projection into the embedding style space (similar to the one used in Jia et al. (2018)). 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.
2. Unlike the work of Ahuja et al. (2020) and previous works, the encoding of the style takes into account 3 modalities: body poses, text semantics, and speech - Mel spectrograms; which are important for gesture generation (Kucherenko et al. (2019); Ginosar et al. (2019a)) and linked to style. We encode and disentangle *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 pose modalities. A fader loss is introduced to effectively disentangle content and style encodings (Lample et al. (2017)).

In the following sections, we first discuss the related works and more specifically the existing behavior style modelling approaches, as well as their limitations. Next, in Section 3, we dive into the details of our model's architecture, describe its training regime, and the objective and subjective evaluations we conducted. We then discuss in Section 4 the objective and subjective evaluation results. Next, in Section 5, we review the key findings of our work, compare it to prior research and discuss its main limitations. We conclude by discussing future directions for our work.

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 (Sargin et al. (2008)), and body motion (Levine et al. (2009); Marsella et al. (2013)).

Chiu and Marsella (2014) 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. LSTM networks driven by speech were recently used to predict sequences of gestures (Hasegawa et al. (2018)) and body motions (Shlizerman et al. (2018); Ahuja et al. (2019)). LSTMs were additionally employed for synthesizing sequences of facial gestures driven by text and speech, namely the fundamental frequency (F0) (Fares (2020); Fares et al. (2021a)). Generative adversarial networks (GANs) were proposed to generate realistic head motion (Sadoughi and Busso (2018)) and body motions (Ferstl et al. (2019)). Furthermore, transformer networks and attention mechanisms were recently used for upper-facial gesture synthesis based on multimodal data - text and speech (Fares et al. (2021b)). Jonell et al. (2020) propose a probabilistic approach based on normalizing flows for synthesizing facial gestures in dyadic settings. Facial (Fares et al. (2021b); Fares (2020)) and hand (Kucherenko et al. (2020)) gestures driven by both acoustic and semantic information 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 (Neff et al. (2008); Karras et al. (2017); Cudeiro et al. (2019); Ahuja et al. (2020); Ginosar et al. (2019a); Alexanderson et al. (2020); Ahuja et al. (2022)). Neff et al. (2008) propose a system that produces full body gesture animation driven by text, in the style of a specific performer. Alexanderson et al. (2020) propose a generative model for synthesizing speech-driven gesticulation, they exert directorial control over the output style such as gesture level and speed. Karras et al. (2017) 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. Cudeiro et al. (2019) also propose a model that synthesizes 3D facial animation driven by speech signal. The learned model, VOCA (Voice Operated Character Animation) takes any speech signal as input–even speech in languages other than English–and realistically animates a wide range of adult faces. Conditioning on subject labels during training allows the model to learn a variety of realistic speaking styles. VOCA also provides animator controls to alter speaking style, identity-dependent facial shape, and pose (i.e. head, jaw, and eyeball rotations) during animation.

Ginosar et al. (2019a) 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 (Neff et al. (2008); Karras et al. (2017); Cudeiro et al. (2019); Ahuja et al. (2020)). 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 Ahuja et al. (2020) and Ahuja et al. (2022). Ahuja et al. (2020) 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. Moreover, the style transfer is limited to the styles of PATS speakers, 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, Ahuja et al. (2022) 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 MATERIALS AND METHODS

3.1 Model Architecture

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 that have been seen and unseen during training.

As depicted in Figure 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 upper-body behavior (including hand gestures and body poses) 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 (Lample et al. (2017)) 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. Gesture styles for the same input speech can be directly

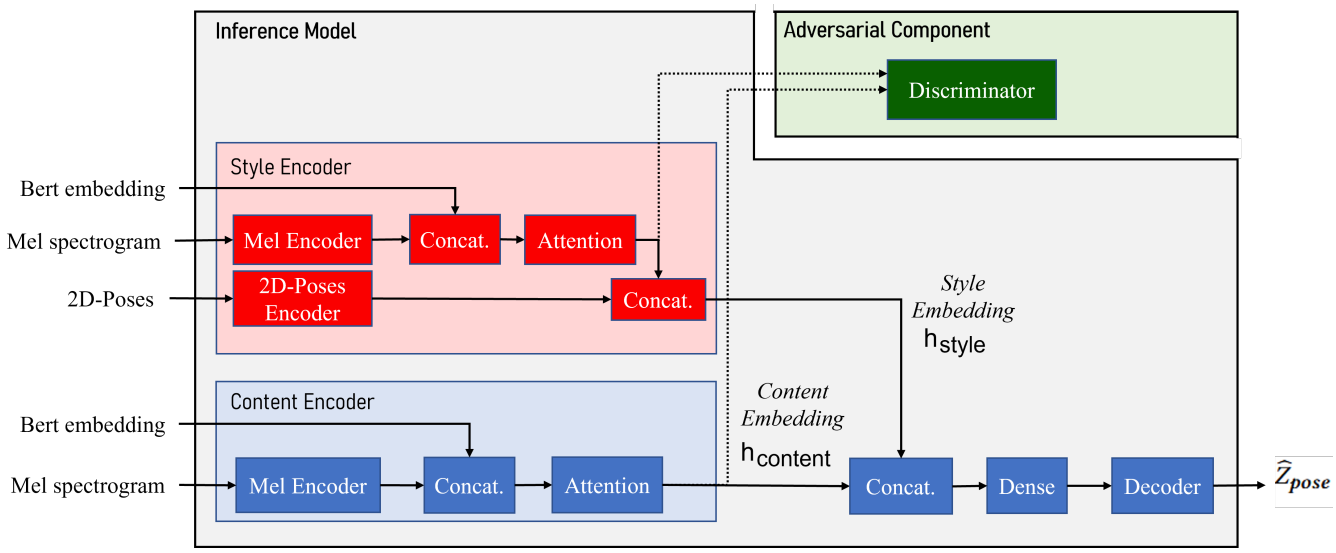

Figure 1. ZS-MSTM (**Z**ero-**S**hot **M**ultimodal **S**tyle **T**ransfer **M**odel) architecture. The content encoder (further referred to as $E_{content}$) is used to encode content embedding $h_{content}$ from BERT text embeddings X_{text} and speech Mel-spectrograms X_{speech} using a speech encoder $E^{content}_{speech}$. The style encoder (further referred to as E_{style}) is used to encode style embedding h_{style} from multimodal text X_{text}, speech X_{speech}, and pose X_{pose} using speech encoder E^{style}_{speech} and pose encoder E^{style}_{pose}. 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 Dis is used to disentangle content and style embeddings $h_{content}$ and h_{style}.

controlled by switching the value of the style embedding vector h_{style} or by calculating this embedding from a target speaker's multimodal 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 a segment-level of 64 timesteps: the inputs and output of the network consist of one feature vector for each segment ***S*** of the input text sequence. The length of the segment-level input features (text and audio) corresponds to $t = 64$ timesteps (as provided by *PATS Corpus*). The model generates a sequence of gestures corresponding to the same segment-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.

3.1.1 Content Encoder

The content encoder $E_{content}$ illustrated in Figure 1 takes as inputs BERT embedding X_{text} and audio Mel spectrograms X_{speech} corresponding to each **S**. X_{text} is represented by a vector of length 768 - BERT embedding size used in *PATS Corpus*. X_{speech} is encoded using *Mel Spectrogram Transformer (AST)* pre-trained *base384* model (Gong et al. (2021)).

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 **S**. The *Transformer Encoder* has an embedding dimension equals to d_{model}, N_{enc} equals to 12 encoding layers, and N_h equals to 12 attention heads.

The segment-level encoded Mel spectrogram is then concatenated with the segment-level BERT embedding. A self-attention mechanism is then applied on the resulting vector. The multi-head attention layer has N_h equals to 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 segment-level Mel spectrogram and text embedding, and it can be written as follows:

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

where: sa(.) denotes self-attention.

3.1.2 Style Encoder

As discussed previously, *behavior 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 *behavior style* is encoded in a speaker's multimodal - text, speech and pose - behavior. As illustrated in Figure 1, the style encoder E_{style} takes as input, at the segment-level, Mel spectrogram X_{speech}, BERT embedding X_{text}, and a sequence of (X, Y) joints positions that correspond to a target speaker's 2D poses X_{pose}. *AST* is used to encode the audio input spectrogram. N_{lay} equals to 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 N_h equals to 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:

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

where: sa(.) denotes self-attention.

3.1.3 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 S-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 segment 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_{dec} = *1* decoding layer, with N_h = 2 attention heads, and an embedding size equal to d_{model}. Similar to the one proposed in Vaswani et al. (2017), 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 ***j*** depends only on the known outputs at positions that are less than ***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:

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

3.1.4 Adversarial Component

Our approach of disentangling style from content relies on the fader network disentangling approach (Lample et al. (2017)), where a fader loss is introduced to effectively separate content and style encodings, as depicted in Figure 2. 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 *Dis* 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 *Dis* 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:

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

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

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

In turn, optimizing the generator loss including the fader loss $\mathcal{L}^{gen}_{adv}$, 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,

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

To be consistent, the style prediction error is preliminary normalized within 0 and 1 range.

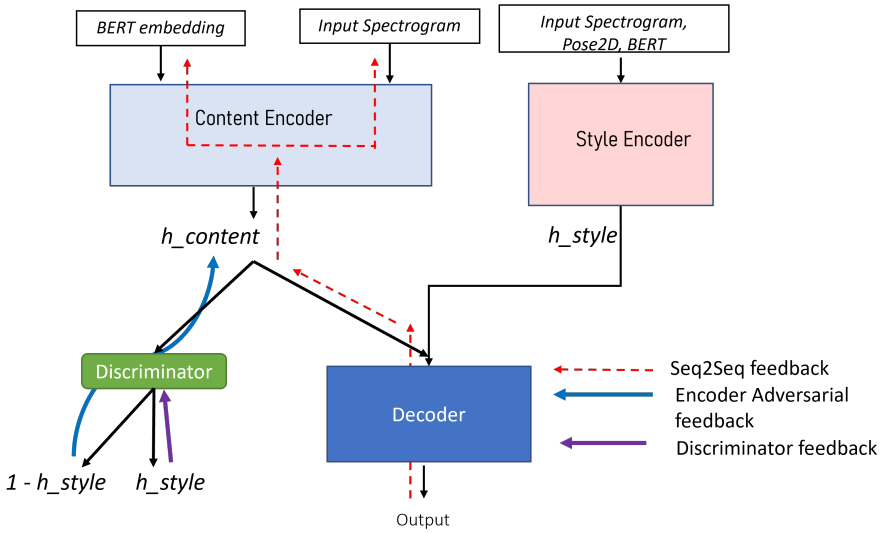

Figure 2. Fader network for multimodal content and style disentangling.

Finally, the total generator loss can therefore be written as follows:

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

where λ is the adversarial weight that starts off at 0 and is linearly incremented by 0.01 after each training step.

The discriminator Dis and the generator G are then optimized alternatively as described in Lample et al. (2017).

All **ZS-MSTM** hyperparameters were chosen empirically and are summarized in Table 2.

3.2 Training Regime

This section describes the training regime we follow for training **ZS-MSTM**. We trained our network using the *PATS Corpus* (Ahuja et al. (2020)). 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 model 11 joints that represent upper body and arm joints.

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 Ahuja et al. (2020).

For each PATS speaker, there is a train, validation and test set already defined in the database. For testing the *Seen Speaker* condition, our test set includes the train sets of 16 PATS speakers. Six other speakers are selected for the *Unseen Speaker* condition, and their test sets are also used for our experiments. These six speakers differ in their behavior style and lexical diversity. *Seen* and *Unseen* speakers are listed in Table 3.

We developed our model using Pytorch and trained it on an NVIDIA Corporation GP102 (GeForce GTX 1080 Ti) machine. Each training batch contains BS = 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 Lr equals to 1e-5, with W_{steps} equals to $20,000$. We train the network for N_{ep} = 200. All features values are normalized so that the dataset mean and standard deviation are 0 and 0.5, respectively. Table 4 summarizes all hyperparameters used for training.

3.3 Objective Evaluation

To validate our approach and assess the stylized generated gestures, we conduct an objective evaluation for the two conditions *Seen Speakers* and *Unseen Speakers*.

3.3.1 Objective Metrics

In our work, we have defined *behavior style* by the *behavior expressivity* of a speaker. To evaluate objectively our works, we define metrics to compare the *behavior expressivity* generated by our model, with the target speaker's *behavior expressivity*, and source speaker's *behavior expressivity*.

Following works on *behavior expressivity* by Wallbott (1998) and Pelachaud (2009), 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* (BB 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* (Wallbott (1998); Kucherenko et al. (2019)).

For both conditions SD and SI, we define two sets of distances:

1. ***Dist.****(Source, Target)*: representing the average distance between the source style and the target style,
2. ***Dist.****(ZS-MSTM, Target)*: representing the average distance between our model's gestures style and the target style.

More specifically, after computing the *behavior expressivity* and *BB perimeter* of our model's generated gestures, the ones of source speakers, and the ones of the target speakers, we calculate the average distance as follows:

$$\textbf{Dist}_{avg}\textit{(x, Target)} = \frac{\textbf{Dist.}\textit{(x, Target)}}{\textbf{Dist.}\textit{(Source, Target)} + \textbf{Dist.}\textit{(ZS-MSTM, Target)}} \times 100 \quad (9)$$

Where x denotes *Source* for computing $\textbf{Dist}_{avg}(Source, Target)$ and *ZS-MSTM* for computing $\textbf{Dist}_{avg}(\textit{ZS-MSTM}, Target)$.

To investigate the impact of each input modality on our style encoder, we conducted ablation studies on different versions of our model. Specifically, we performed ablations of the *pose modality*, *text modality*, and *audio modality*. We also compared the performance of the full model with that of the baseline *DiffGAN* Ahuja et al. (2022). We employ two metrics to evaluate the correlation and timing between gestures and spoken language: **Probability of Correct Keypoints (PCK)** and **L1 distance**. For PCK, we averaged the values over $\alpha = 0.1$ and 0.2, as suggested in Ginosar et al. (2019b). L1 distance was calculated between the generated gestures and the corresponding target ground truth gestures.

3.4 Human Perceptual Studies

We conduct three human perceptual studies.

1. **Study 1** - To investigate human perception of the stylized upper-body gestures produced by our model, we conduct a human perceptual study that aims to assess the style transfer of speakers *seen* during training - *Seen Speaker* condition.
2. **Study 2** - We conduct another human perceptual study that aims to assess the style transfer of speakers *unseen* during training - *Unseen Speaker* condition.
3. **Study 3** - We additionally conduct a third human perceptual study to compare **ZS-MSTM**'s produced stylized gestures in *Seen Speaker* and *Unseen Speaker* conditions, to *Mix-StAGE* which we consider our baseline.

The evaluation studies are conducted with 35 participants that were recruited through the online crowd-sourcing website Prolific. Participants are selected such that they are fluent in English. 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 in the form of 2D sticks.

Study 1 and 2. For Study 1 and 2, we presented 60 stimuli of 2D stick animations. Each study included 30 stimuli. A stimulus is a triplet of 2D animations composed of:

- A 2D animation with the *source style*,
- A 2D animation with the *target style*,
- A 2D animation of *ZS-MSTM*'s prediction after performing the style transfer.

Figure 3 illustrates the three animations we present for each set of questions. The animation of the target style is the ***Reference***. The animation of our model's predictions, and the source style is either ***Animation A*** or ***Animation B*** (randomly chosen).

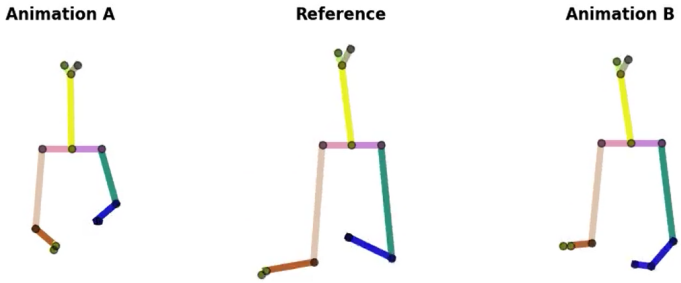

Figure 3. Three 2D stick animations: *Animation A*, the *Reference*, and *Animation B*. The target style is represented by *Reference*. **ZS-MSTM**'s predictions, and the **source style** are illustrated in Animation A or B.

For each triplet of animations, we asked 6 questions to evaluate 6 factors related to the *resemblance* of the produced gestures w.r.t the the *source style* and *target style*:

1. Please rate **the overall resemblance of the Reference** w.r.t A and B. (**Factor 1** - Overall resemblance)
2. Please rate the **resemblance of the Left (L) and Right (R) arms gesturing of the Reference** w.r.t the left and right arm gesturing of A and B. (**Factor 2** - Arms gesturing)
3. Please rate the **resemblance of the body orientation of the Reference** w.r.t the body orientation of A and B. (**Factor 3** - Body orientation)
4. Please rate the **resemblance of the gesture amplitude of the Reference** w.r.t the gesture amplitude of A and B. (**Factor 4** - Gesture amplitude)
5. Please rate the **resemblance of the gesture frequency of the Reference** w.r.t the gesture frequency of A and B. (**Factor 5** - Gesture frequency)
6. Please rate the **resemblance of the gesture velocity of Reference** w.r.t the gesture velocity of A and B. (**Factor 6** - Gesture velocity)

Each factor is rated on a *5 likert* scale, as follows:

1. Reference is very similar to A
2. Reference is mostly similar to A
3. Reference is in between A and B
4. Reference is mostly similar to B
5. Reference is very similar to B

Training. Each study includes a training at its beginning. The training provides an overview of the 2D upper-body skeleton of the virtual agent, its composition, and gesturing. The goal of the training is to get the participants familiarized with the 2D skeleton before starting the study. More specifically, the training included a description of how the motion of a speaker in a video is extracted by detecting his/her facial and body motion and extracting his/her 2D skeleton of joints, and stated that in a similar fashion, the eyes and upper-body movement of a virtual agent are represented by a 2D skeleton of joints, as depicted in Figure 4

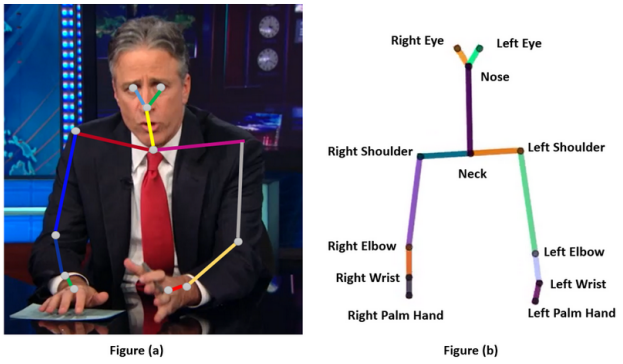

Figure 4. Upper-body 2D skeleton of a speaker Vs. a virtual agent

Moreover, we present and describe different shots of the 2D skeleton gesturing with *its right/left arms*, and with different *body orientation*, which is described as the orientation of the shoulders and neck.

Pre-tests. We conducted pre-tests to make sure that the 2D animations are comprehensible by participants, as well as the questions. Participants reported that the training, stimuli and questions are coherent and comprehensible, however each study was too long, as it lasted 30 minutes. For this reason, we divided each study to three, such that each study includes only 10 stimuli, and is conducted by different participants. Hence, 6 studies including a pre-training, and the evaluation of 10 stimuli were conducted by 35 participants that are different.

Study 3. For Study 3, we present 20 stimuli consisting of triplets of 2D stick animations. Similar to *Study 1* and *Study 2*, for each triplet, we present: *Animation A*, the *Reference*, and *Animation B*. The animation of the target style is the *Reference*. The animation of Mix-StAGE's predictions, and the source style is either Animation A or Animation B (randomly chosen). We note that these stimuli include the same *source* and *target* styles that were used in *Study 1* and *Study 2*, and which were randomly chosen. Study 3 also included training at its beginning, which is the same as the one previously described.

4 RESULTS

4.1 Objective Evaluation Results

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.

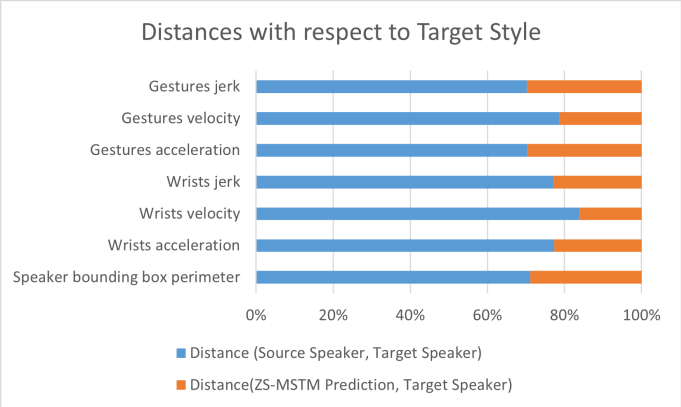

Figure 5. Distances between the target speaker style and each of the source style and our model's generated gestures style for seen target speakers

Figure 5 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.

For *Seen Speaker* condition (Figure 5), ***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 (Ahuja et al. (2020)).

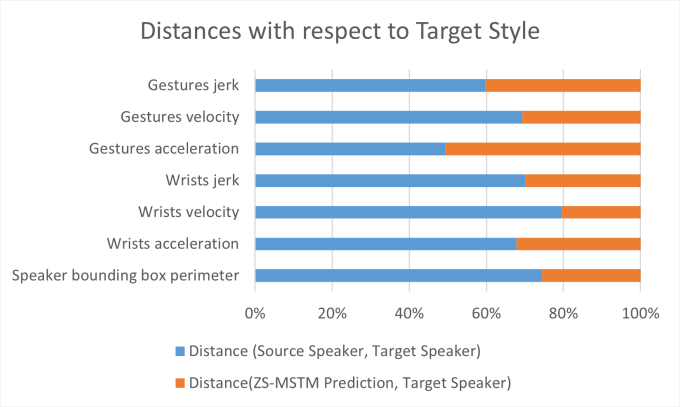

Figure 6. Distances between the target speaker style and each of the source style and our model's generated gestures style for unseen target speakers

Experimental results for the *Unseen Speaker* test set are depicted in Fig. 6. Results reveal that our model is capable of reproducing the style of the 6 unseen speakers. As depicted in Fig. 6, 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.

The results of our ablation studies are summarized in Table 1. Specifically, we trained three versions of our *ZS-MSTM* model, each with one modality (either text, audio, or pose) removed from the style encoder. We evaluated the performance of each model using the *L1 distance* and *PCK* metrics, comparing

the predictions to the target ground truth in all conditions. Our results (see Table 1) show that the *L1 distance* between the predictions of the ablated models and the ground truth is higher compared to the full model condition, for both seen (*Oliver*) and unseen (*Chemistry, Maher*) target styles. This trend was observed across all three ablation conditions. In addition, we compared our results to the baseline *DiffGAN* (Ahuja et al. (2022)) and found that our ***ZS-MSTM*** model consistently outperforms *DiffGAN* in terms of *L1 distance*, with higher confidence intervals reported as standard deviation on all source-target pairs. Furthermore, we evaluated the ***PCK*** metric for all source-target pairs, and found that our ***ZS-MSTM*** model achieves higher accuracy than the ablated models for all style transfers, with higher confidence intervals. This indicates that our model produces joint positions that are accurate and closely match the ground truth. When comparing ***ZS-MSTM*** with *DiffGAN*, our model outperforms *DiffGAN* in terms of *PCK*, with higher confidence intervals.

4.1.1 Additional t-SNE Analysis

In this work, the style encoder is agnostic: it is the attention weights that make it possible to exploit the different modalities given as input to the style encoder.

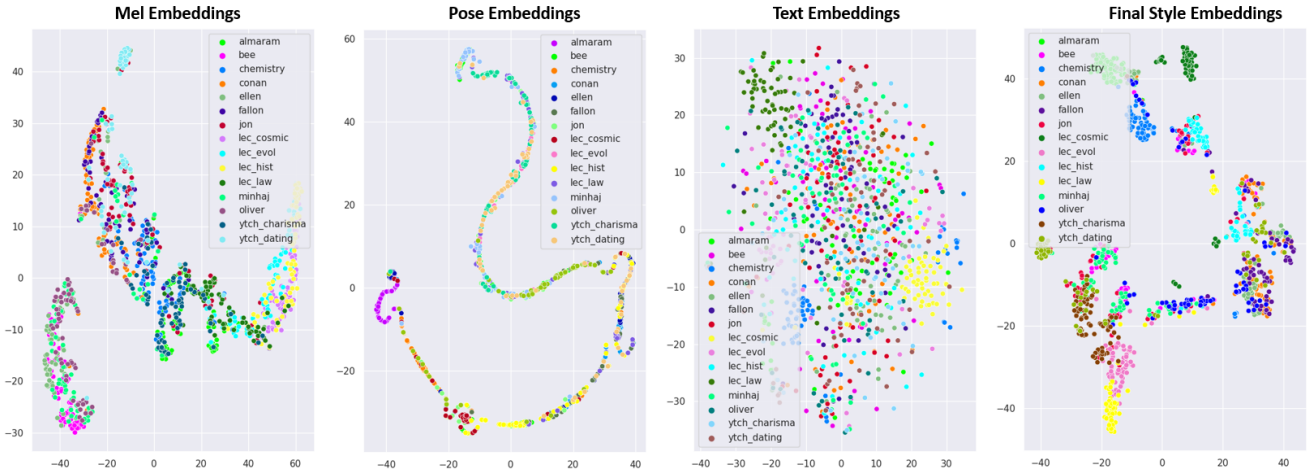

Figure 7. 2D TSNE Analysis of the generated ***Mel Embeddings***, ***Pose Embeddings***, ***Text Embeddings***, and the final ***Style Embeddings***

We conducted a t-SNE post-hoc analysis of the distributions of the style vectors at the output of each modality. Figure 7 illustrates the 2D t-SNE plots of ***Mel Embeddings***, ***Pose Embeddings***, ***Text Embeddings***, and the final ***Style Embeddings*** produced by our model **ZS-MSTM**. We found that the motion style depends most on the *pose modality*, followed by the *speech*, then the *text semantics*.

4.2 Human Perceptual Studies Results

Study 1 - *Seen Speakers*.

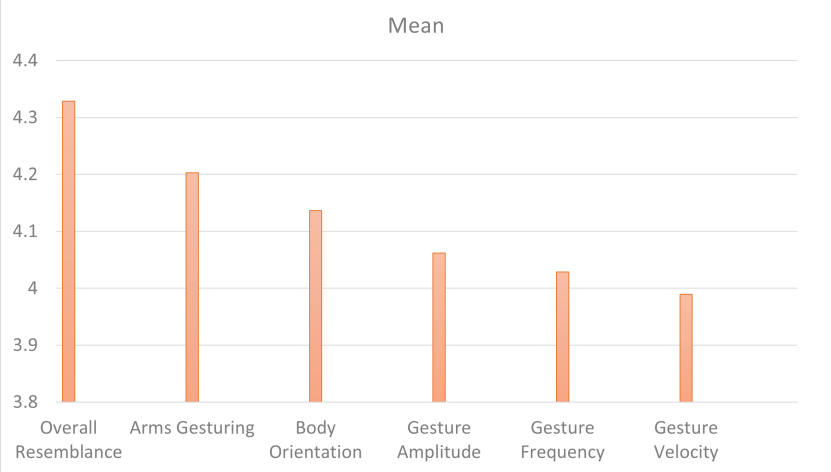

Figure 8. The mean scores of all the factors for *Seen Speakers* condition

Our first perceptive study (Study 1) aims to evaluate the style transfer of speakers *seen* during training. Figure 8 shows the mean scores obtained on the 6 factors for the condition *"seen speakers"*. On a *5 likert scale*, the ***overall resemblance*** factor obtained a score of 4.32, which means that the **ZS-MSTM**'s 2D animations closely resemble the 2D animations of the *seen target style*. The resemblance is also reflected by the mean scores of ***arms gesturing***, ***body orientation***, ***gesture amplitude***, ***gesture frequency***, as well as ***gesture velocity***, which is between 3.99 and 4.2. We observed that for all factors, most of the participants gave a score between 3.8 and 5, as depicted in Figure 9.

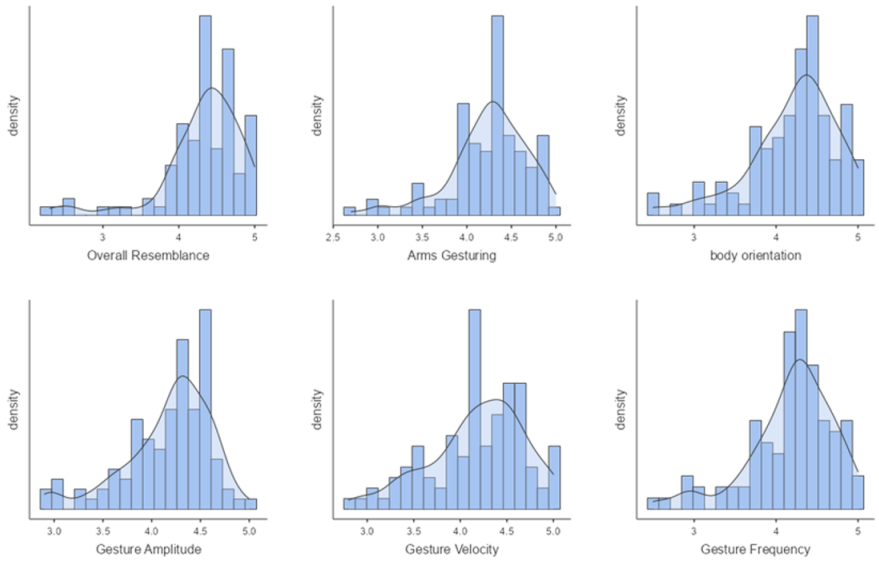

Figure 9. Density plots of ***Overall Resemblance***, ***Arms Gesturing***, ***Body Orientation***, ***Gesture Amplitude***, ***Gesture Frequency***, ***Gesture Velocity*** for the *Seen Speakers* condition

We additionally performed post-hoc paired samples t-tests between all the factors, and found significant results between ***overall resemblance*** and all the other factors ($p \leq 0.008$).

Study 2 - *Unseen Speakers.*

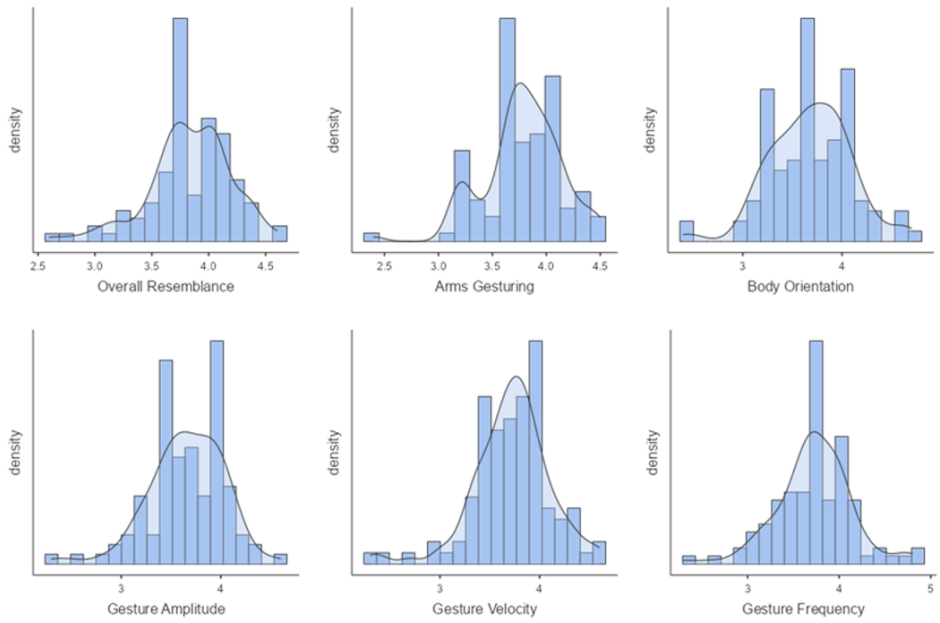

Figure 11. ***Body Orientation***, ***Gesture Amplitude***, ***Gesture Frequency***, ***Gesture Velocity*** for the *Unseen Speakers* condition

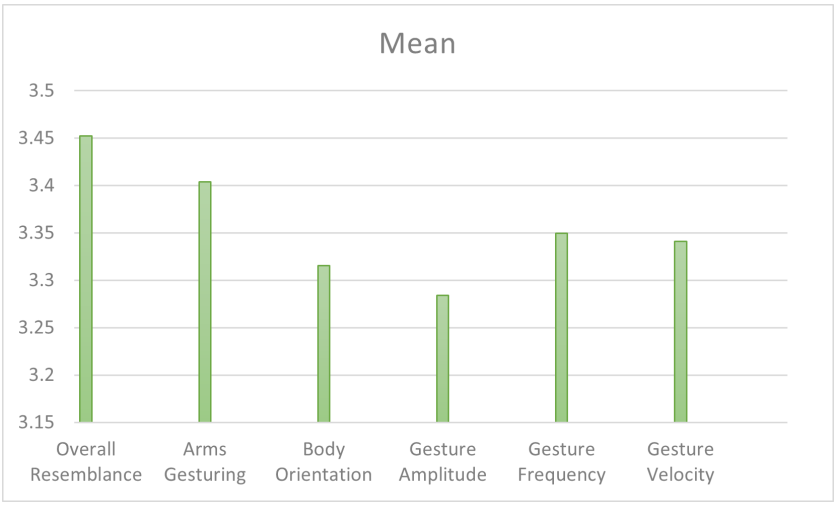

Figure 10. The mean scores of all the factors for *Unseen Speakers* condition

Our second perceptive study (Study 2) aims to evaluate the style transfer of speakers *unseen* during training. Figure 10 illustrates the mean scores obtained on the 6 factors for the condition *"unseen speakers"*. On a *5 likert scale*, the ***overall resemblance*** factor obtained a score of 3.45, which means that there is an overall resemblance between **ZS-MSTM**'s 2D animations and the *unseen target style*. The resemblance is also reflected by the mean scores of ***arms gesturing***, ***body orientation***, ***gesture amplitude***, ***gesture frequency***, as well as ***gesture velocity***, which is between 3.28 and 3.41. We observed that for all factors, most of the participants gave a score between 3 and 4, as depicted in Figure 11.

We additionally performed post-hoc paired samples t-tests between all the factors, and found significant results between ***overall resemblance*** and all the other factors ($p \leq 0.014$).

Study 3 - Comparing with Mix-StAGE. The third perceptive study aims to compare the performance of our model with respect to the State of the Art, *Mix-StAGE*. Figure 12 illustrates the mean scores obtained for the two conditions *Mix-StAGE* and *ZS-MSTM*, w.r.t the 6 factors. As shown in Figure 12, for all the factors,

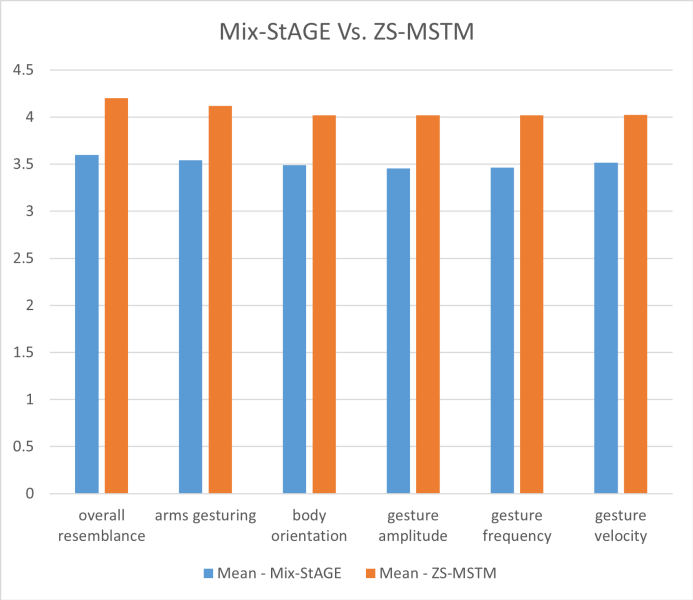

Figure 12. ***ZS-MSTM*** Vs. ***Mix-StAGE***

our model obtained higher mean scores than *Mix-StAGE*. Our model performs better than *Mix-StAGE* in terms of the ***overall resemblance*** of the generated gestures w.r.t the animations produced with the *target style* (mean score *ZS-MSTM* (4.2) $\geq$ mean score *Mix-StAGE* (3.6)). More specifically, the resemblance between the synthesized 2D gestures of *ZS-MSTM* and the target style is greater than the one between *Mix-StAGE* and the target style. This result is also reflected in the resemblance of the ***arms gesturing***, ***body orientation***, ***gesture amplitude***, ***gesture frequency*** and ***gesture velocity*** of our model's produced gestures w.r.t the *target style*. More specifically, our model obtained a mean score between 4 and 4.2 for all the factors, while *Mix-StAGE* obtained a mean score between 3.8 and 3.6 for all the factors. We additionally conducted post-hoc paired t-tests between the factors in condition *Mix-StAGE* and those in *ZS-MSTM*. We found significant results between all the factors in the condition *Mix-StAGE* and those in *ZS-MSTM* ($p < 0.001$ for all). These results show that the mean scores for all the factors in condition *ZS-MSTM* are significantly greater than those *Mix-StAGE*. Thus, we can conclude that our model *ZS-MSTM* can successfully render animations with the style of another speaker, going beyond the state of the art *Mix-StAGE*.

5 DISCUSSION AND CONCLUSION

We have presented ***ZS-MSTM***, the first approach for zero-shot multimodal style transfer for 2D pose synthesis that allows the transfer of style from any speakers *seen* or *unseen* during the training phase. To the best of our knowledge, our approach ***ZS-MSTM*** 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. ***ZS-MSTM*** can learn the style latent space of speakers, given their multimodal data, and independently from their identity. It can synthesize body gestures of a source speaker, given the source

speaker's mel spectrogram and text semantics, with the style of another target speaker given the target speaker's multimodal behavior style that is encoded through the mel spectrogram, text semantics, and pose modalities. Moreover, our approach is *zero-shot*, thus is capable of transferring the style of unseen speakers. It is not not limited to *PATS* speakers, and can produce gesture in the style of any newly coming speaker without further training or fine-tuning, rendering our approaches *zero-shot*. *Behavioral style* is modelled based on multimodal speakers' data, and is *independent* from the *speaker's identity* ("ID"), which allows our model to generalize style to new *unseen* speakers. We validated our approach by conducting objective and subjective evaluations. The results of these studies showed that ***ZS-MSTM*** generates stylized animations that are close to the target style, for target speakers that are *seen* and *unseen* by our model. The results of our ablation studies (see Table 1) suggest that all three modalities (text, audio, and pose) are important for the performance of our *ZS-MSTM* model in style transfer tasks. When any one of these modalities is removed from the style encoder, the *L1 distance* between the model's predictions and the ground truth increases, indicating lower performance. This shows the importance of incorporating multiple modalities for better style transfer in our model. Moreover, we compared the performance of ***ZS-MSTM*** w.r.t the state of the art ***Mix-StAGE*** and results showed that ***ZS-MSTM*** performs better in terms of *overall resemblance* of the generated gestures w.r.t the animations produced with the *target style*. ***ZS-MSTM*** can generalize style to new speakers without any fine-tuning or additional training, unlike *Mix-StAGE*. Its independence from the speaker's identity "ID" allows the generalization without being constrained and limited to the speakers used for training the model. DiffGAN was later on proposed by Ahuja et al. (2022) as an extension to *Mix-StAGE*, and an approach that performs *few-shot* style transfer strategy based on neural domain adaptation accounting for cross-modal grounding shift between source speaker and target style. However this adaptation still requires 2 minutes of the style to be transferred which is not required by our model. Our comparison with the baseline *DiffGAN* model shows that our ***ZS-MSTM*** model outperforms it in terms of both *L1 distance* and *PCK* metrics. This shows that our model is better at generating accurate human poses, especially when transferring styles that it has not seen during training. Overall, our results suggest that our *ZS-MSTM* model is a promising approach for style transfer tasks in human pose estimation, as it can leverage multiple modalities to generate poses that are accurate.

Our approach allows the transfer of style from any speakers *seen* or *unseen* during the training phase. *behavior style* was never viewed as being *multimodal*; previous works limit behavior style to arm gestures only. However, both *text* and *speech* convey *style* information, and the embedding vector of *style* must consider the three modalities. Our assumption was confirmed by our post-hoc t-SNE analysis of the distributions of the style vectors at the output of each modality. We found that the motion style depends mainly on the body *pose modality*, followed by the *speech modality*, then the *text semantics modality*. We conducted an objective evaluation and three perceptive studies. The results of these studies show that our model produces stylized animations that are close to the target speakers style even for *unseen* speakers.

While we have made some strides, there are still some limitations. The main limitation of ***ZS-MSTM*** is that it was not evaluated on an ECAs. The main reason is that it was trained on the *PATS Corpus*, which include 2D poses. The graphical representation of the data as 2D stick figure is not always readable, even when being projected on the video of a human speaker. The main reason behind this problem is that the animation is missing information on the body pose in the Z direction (the depth axis). An interesting direction for future work is to extend our model to capture the different gesture shapes and motion. Gesture shapes convey different meanings. For example, a pointing index can indicate a direction. Hand shapes and arm movement can describe an object, an action, etc. Several attempts have looked at modelling metaphoric gestures (Ravenet et al. (2018)), or iconic gestures (Bergmann and Kopp (2009)). Most generative models of gestures do not compute the gesture shapes and motions for those specific gesture types. Extending our

model to capture gesture shapes and motion would require extending the Corpora PATS, to include specific annotations related to gestures shapes and to identify better representations (such as image schemas Grady (2005) for metaphoric gestures).

6 ACKNOWLEDGMENT

This work was performed within the Labex SMART (ANR-11-LABX-65) supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02. Part of this work - a short description of ZS-MSTM, and of the objective evaluation and results - will appear in the proceedings of SIVA'23 the Workshop on Socially Interactive Human-like Virtual Agents, satellite of the conference face and gesture FG'23.

REFERENCES

Ahuja, C., Lee, D. W., and Morency, L.-P. (2022). Low-resource adaptation for personalized co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*

Ahuja, C., Lee, D. W., Nakano, Y. I., and Morency, L.-P. (2020). Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *European Conference on Computer Vision* (Springer), 248–265

Ahuja, C., Ma, S., Morency, L.-P., and Sheikh, Y. (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

Alexanderson, S., Henter, G. E., Kucherenko, T., and Beskow, J. (2020). Style-controllable speech-driven gesture synthesis using normalising flows. In *Computer Graphics Forum* (Wiley Online Library), vol. 39, 487–496

Bell, A. (1984). Language style as audience design. *Language in society* 13, 145–204

Bergmann, K. and Kopp, S. (2009). Gnetic–using bayesian decision networks for iconic gesture generation. In *International workshop on intelligent virtual agents* (Springer), 76–89

Campbell-Kibler, K., Eckert, P., Mendoza-Denton, N., and Moore, E. (2006). The elements of style. In *Poster presented at New Ways of Analyzing Variation*. vol. 35

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

Chiu, C.-C. and Marsella, S. (2014). Gesture generation with low-dimensional embeddings. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 781–788

Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., and Black, M. J. (2019). Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10101–10111

Fares, M. (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

Fares, M., Pelachaud, C., and Obin, N. (2021a). Multimodal-based upper facial gestures synthesis for engaging virtual agents. In *WACAI 2021*

Fares, M., Pelachaud, C., and Obin, N. (2021b). Multimodal generation of upper-facial and head gestures with a transformer network using speech and text. *arXiv preprint arXiv:2110.04527*

Ferstl, Y., Neff, M., and McDonnell, R. (2019). Multi-objective adversarial gesture generation. In *Motion, Interaction and Games*. 1–10

Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., and Malik, J. (2019a). Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*

Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., and Malik, J. (2019b). Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3497–3506

Gong, Y., Chung, Y.-A., and Glass, J. (2021). Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*

Grady, J. E. (2005). Image schemas and perception: Refining a definition. *From perception to meaning: Image schemas in cognitive linguistics* 29, 35

Hasegawa, D., Kaneko, N., Shirakawa, S., Sakuta, H., and Sumi, K. (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

Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., et al. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems* 31

Jonell, P., Kucherenko, T., Henter, G. E., and Beskow, J. (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

Karras, T., Aila, T., Laine, S., Herva, A., and Lehtinen, J. (2017). Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)* 36, 1–12

Kucherenko, T., Hasegawa, D., Henter, G. E., Kaneko, N., and Kjellström, H. (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

Kucherenko, T., Jonell, P., van Waveren, S., Henter, G. E., Alexanderson, S., Leite, I., et al. (2020). Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the ACM International Conference on Multimodal Interaction*

Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., and Ranzato, M. (2017). Fader networks: Manipulating images by sliding attributes. *Advances in neural information processing systems* 30

Levine, S., Theobalt, C., and Koltun, V. (2009). Real-time prosody-driven synthesis of body language. In *ACM SIGGRAPH Asia*. 1–10

Lugrin, B. (2021). Introduction to socially interactive agents. In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*. 1–20

Marsella, S., Shapiro, A., Feng, A., Xu, Y., Lhommet, M., and Scherer, S. (2013). Towards higher quality character performance in previz. In *Proceedings of the Symposium on Digital Production*. 31–35

McNeill, D., Bertenthal, B., Cole, J., and Gallagher, S. (2005). Gesture-first, but no gestures? *Behavioral and Brain Sciences* 28, 138–139

Mendoza-Denton, N. (1999). Style. *Journal of Linguistic Anthropology* 9, 238–240

Moon, S., Kim, S., and Choi, Y.-H. (2022). Mist-tacotron: End-to-end emotional speech synthesis using mel-spectrogram image style transfer. *IEEE Access* 10, 25455–25463

Neff, M., Kipp, M., Albrecht, I., and Seidel, H.-P. (2008). Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics (TOG)* 27, 1–24

Norris, S. (2004). *Analyzing multimodal interaction: A methodological framework* (Routledge)

Obermeier, C., Kelly, S. D., and Gunter, T. C. (2015). A speaker's gesture style can affect language comprehension: Erp evidence from gesture-speech integration. *Social cognitive and affective neuroscience* 10, 1236–1243

Obin, N. (2011). *MeLos: Analysis and modelling of speech prosody and speaking style*. Ph.D. thesis

Pelachaud, C. (2009). Studies on gesture expressivity for a virtual agent. *Speech Communication* 51, 630–639

Ravenet, B., Pelachaud, C., Clavel, C., and Marsella, S. (2018). Automating the production of communicative gestures in embodied characters. *Frontiers in psychology* 9, 1144

Sadoughi, N. and Busso, C. (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

Sargin, M. E., Yemez, Y., Erzin, E., and Tekalp, A. M. (2008). Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 1330–1345

Shlizerman, E., Dery, L., Schoen, H., and Kemelmacher-Shlizerman, I. (2018). Audio to body dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7574–7583

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*

[Dataset] Wagner, P., Malisz, Z., and Kopp, S. (2014). Gesture and speech in interaction: An overview

Wallbott, H. (1998). Bodily expression of emotion. *European Journal of Social Psychology* 28, 879–896

7 TABLES

Table 1. Comparison of our ZS-MSTM model with DiffGAN Ahuja et al. (2022) used as a baseline, as well as with different versions of our model where we removed the Text, Audio, and Pose modalities from the Style Encoder. Please note that we report here the values for *DiffGAN* from Ahuja et al. (2022)

.

	L1				PCK			
Models — *Source:* *Target:*	**Oliver Chemistry**	**Maher Chemistry**	**Oliver Maher**	**Maher Oliver**	**Oliver Chemistry**	**Maher Chemistry**	**Oliver Maher**	**Maher Oliver**
ZS-MSTM - Text Ablation	0.51 ± 0.05	0.58 ± 0.07	0.56 ± 0.07	0.36 ± 0.08	0.89 ± 0.78	0.95 ± 0.98	0.87 ± 0.89	0.97 ± 0.76
ZS-MSTM - Audio Ablation	0.65 ± 0.08	0.71 ± 0.08	0.91 ± 0.07	0.89 ± 0.08	0.85 ± 0.78	0.82 ± 0.98	0.84 ± 0.89	0.95 ± 0.76
ZS-MSTM - Pose Ablation	0.87 ± 0.08	0.91 ± 0.08	1.11 ± 0.12	0.76 ± 0.08	0.81 ± 0.78	0.9 ± 0.98	0.82 ± 0.89	0.92 ± 0.76
DiffGAN (Ahuja et al. (2022))	1.36 ± 0.03	0.88 ± 0.03	1.48 ± 0.01	0.53 ± 0.02	0.29 ± 0.01	0.31 ± 0.01	0.26 ± 0.01	0.45 ± 0.02
ZS-MSTM - Full Model	**0.34 ± 0.04**	**0.36 ± 0.04**	**0.49 ± 0.05**	**0.11 ± 0.03**	**0.96 ± 0.91**	**0.96 ± 0.99**	**0.89 ± 0.92**	**0.97 ± 0.98**

Table 2. ZS-MSTM hyperparameters

Component	**Hyperparameter**		**Value**
AST (base384 model)	Embedding size	d_{model}	768
	Encoding layers	N_{lay}	12
	Attention heads	N_h	12
Content Encoder	Attention heads	N_h	4
	Embedding size	d_{att}	1536
Style Encoder	2D Pose LSTMs	N_{lay}	3
		N_{hid}	768
	Attention heads	N_h	4
	Embedding size	d_{att}	1536
Sequence to Sequence Component	Transformer Decoder	N_{dec}	1
	Attention heads	N_h	2
	Embedding size	d_{model}	768

Table 3. Seen and Unseen PATS Speakers

Condition	**Speakers**
Seen	”Shelly”, ”Jon”, ”Fallon”, ”Bee”, ”Ellen”, ”Oliver”, ”Lec_cosmic”, ”Lec_hist”, ”Ytch_prof”, ”Ytch_dating”, ”Seth”, ”Conan”, ”Angelica”, ”Rock”, ”Noah”, and ”Lec_law”
Unseen	”Lec_evol”, ”Almaram”, ”Huckabee”, ”Maher” ”Ytch_charisma”, ”Minhaj”, and ”Chemistry

Table 4. Training Hyperparameters

Hyperparameter		**Value**
Batch Size	BS	24
Number of epochs	N_{ep}	200
Adam Optimizer	β_1	0.95
	β_2	0.999
Scheduler	W_{steps}	20,000
	Lr	1e-5