TranSTYLer: Multimodal Behavioral Style Transfer for Facial and Body Gestures Generation

Mireille Fares, Catherine Pelachaud, Nicolas Obin

TRANSTYLER: MULTIMODAL BEHAVIORAL STYLE TRANSFER FOR FACIAL AND BODY GESTURES GENERATION

Mireille Fares
ISIR, STMS
Sorbonne University
Paris, France

Catherine Pelachaud
ISIR, CNRS
Sorbonne University
Paris, France

Nicolas Obin
STMS
Sorbonne University
Paris, France

ABSTRACT

This paper addresses the challenge of transferring the behavior expressivity style of a virtual agent to another one while preserving behaviors shape as they carry communicative meaning. Behavior expressivity style is viewed here as the qualitative properties of behaviors. We propose *TranSTYLer*, a multimodal transformer-based model that synthesizes the multimodal behaviors of a source speaker with the style of a target speaker. We assume that behavior expressivity style is encoded across various modalities of communication, including text, speech, body gestures, and facial expressions. The model employs a style-content disentanglement schema to ensure that the transferred style does not interfere with the meaning conveyed by the source's behaviors. Our approach eliminates the need for style labels and allows the generalization to styles that have not been seen during the training phase. We train our model on the *PATS corpus*, which we extended to include dialog acts and 2D facial landmarks. Objective and subjective evaluations show that our model outperforms state-of-the-art models in style transfer for both seen and unseen styles during training. To tackle the issues of style and content leakage that may arise, we propose a methodology to assess the degree to which behavior and gestures associated with the target style are successfully transferred, while ensuring the preservation of the ones related to the source content.

1 Introduction

Human communication is a complex phenomenon that involves various modes of expression beyond speech production. It is inherently multimodal, as it relies on the interplay of verbal and non-verbal signals to convey semantic and pragmatic content and facilitate the communication process [Knapp et al.(2013), Argyle(2013), Feyereisen et al.(1991), Armstrong et al.(1995)]. *Human behavior expressivity style* refers to the unique and characteristic pattern of behavior exhibited by an individual in various social and communicative contexts [Knapp et al.(2013)]. It is not a fixed attribute of the speaker but rather is constantly adjusted, accomplished, and co-created with the audience [Mendoza-Denton(1999)]. It involves the way an individual communicates verbally and non-verbally, including verbal skills, body language, gestures, and self-expression. Variability in speakers' gesturing is influenced by

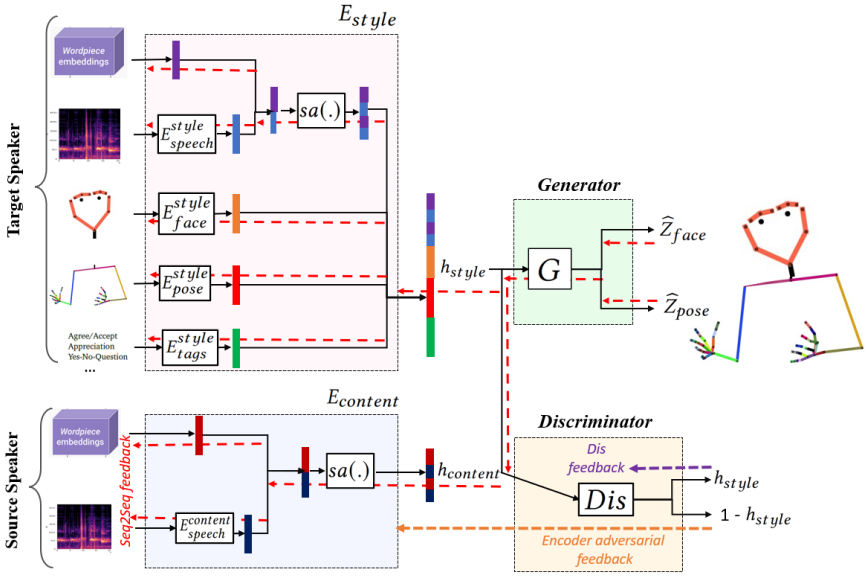


Figure 1: Overview of *TranSTYler*, an approach driven by the content of a source speaker's speech - text semantics (X^{source}_{text}) and Mel spectrogram (X^{source}_{speech}) - and conditioned on a target speaker's style vector h_{style} generated from the target's multimodal input data - Mel spectrogram (X^{target}_{speech}), 2D facial landmarks (X^{target}_{face}), 2D poses (X^{target}_{pose}), and dialog tags (X^{target}_{tags}). The network is composed of: (1) a *style encoder* that encodes the target's multimodal data and generates the style vector h_{style}, (2) a *content encoder* that encodes the source's speech content and generates the content vector $h_{content}$, and (3) a *discriminator* Dis used during training time to disentangle $h_{content}$ and h_{style}. A generator G is used to generate body $\widehat{Z}_{pose}$ and facial $\widehat{Z}_{face}$ gestures.

factors such as their personality traits [McCrae and Costa Jr(1997)], verbal skills [Hostetter and Alibali(2007)], age [Alibali et al.(2009), Feyereisen and Havard(1999)], and culture [Kita(2009)]. The topic and context of the conversation, speaker's role, and relationship with the interlocutor also play a role [Hostetter and Potthoff(2012)]. For example, extroverts tend to use larger spatial gestures[Hostetter and Potthoff(2012)]. Behavior expressivity style can vary between formal and spontaneous speech. In formal settings, a controlled and structured speaking style is used, with formal language and restrained gestures to convey professionalism and respect. In contrast, in spontaneous speech, individuals adopt a more relaxed communication style with informal gestures. In this paper, we propose a novel machine learning approach to synthesize facial and upper-body gestures driven by *prosodic features* and *text* in the style of different speakers including those unseen during training. We view behavior expressivity style as a pervasive factor during speech, influencing the expressiveness of communicative behaviors, while speech content is conveyed through a combination of multimodal behaviors and text. We propose *TranSTYLer* a transformer-based model that can synthesize facial and body gestures of a source speaker in the style of any target speaker, while ensuring that the transferred style does not interfere with the meaning conveyed by the source gestures. Our approach incorporates a disentanglement scheme that separates content and style, enabling us to directly infer the style representation even for speakers who were not part of the training process, without requiring additional training or fine-tuning. Our system comprises two main components. Firstly, we have a speaker style encoder network, which learns to generate a fixed-dimensional embedding that represents the style of a target speaker. This embedding is derived from the target speaker's multimodal data (facial and body gestures, audio, and text). Secondly, we employ a synthesis network that synthesizes gestures based on the content provided by the input modalities (text and audio) of a source speaker. This synthesis process is conditioned on the target speaker style embedding, ensuring that the generated gestures exhibit the target style characteristics. We also introduce a new methodology to measure the transferred style and the preservation of gestures that convey meaning. We evaluate the performance of *TranSTYLer* in terms of *style transfer* and *content preservation*. Objective and subjective evaluations confirm the quality of our approach, outperforming two state-of-the-art models.This paper is organized as follows. We start by providing a review of the existing *behavior expressivity style* modeling approaches, discussing their limitations. We then explain our contributions and the architecture we propose. Finally, we present objective and subjective evaluations and discuss the results.

2 Related Works

Gesture style modeling and control have gained significance in proposing expressive behaviors for virtual agents that can be adjusted and tailored to specific audiences or interlocutors. A first model to generate communicative behaviors with different styles was proposed by Neff et al.[Neff et al.(2008)]. The authors developed a system that generates full body gesture animation based on text, mimicking the style of a specific performer. They focused mainly on hand movements. In recent years, Alexanderson et al. [Alexanderson et al.(2020)] proposed a generative model that synthesizes speech-driven gesticulation while exerting control over the output style, such as gesture level and speed. Karras et al. [Karras et al.(2017)] created a model that generates 3D facial animation from audio to capture the style of a single actor. Similarly, Cudeiro et al. [Cudeiro et al.(2019)] developed a model, called *VOCA*, that synthesizes 3D facial animation driven by speech signal, allowing for a wide range of speaking styles to be realistically animated, even in languages other than English. On the other hand, Ginosar et al. [Ginosar et al.(2019a)] proposed an approach for generating gestures using models trained on individual speakers. Yoon et al [Yoon et al.(2020)] developed a method to generate gestures that matched a speaker's style by using their speaker identity. Their approach involved processing the input audio and text with separate audio and text encoders and using the speaker identity to select a style embedding from a learned space. These features were then fed to a gesture generator to produce a sequence of poses that matched the content and rhythm of the speech. On the other hand, Ghorbani et al. [Ghorbani et al.(2022a), Ghorbani et al.(2022b)] proposed a framework that improved on high-level style portrayal by using exemplar motion sequences to demonstrate the intended stylistic expression of gesture motion. Their method could extract style parameters in a zero-shot manner, only requiring a single example motion and was able to generalize to example motions (and therefore styles) not seen during training. The works just mentioned have focused on generating behaviors from one modality, either facial expressions, head movements or gestures. However, they have not considered multimodal data when modeling style or synthesizing gestures. Additionally, their generative models were only trained on *single-speaker* data. Recently, Ahuja et al. [Ahuja et al.(2020)] have attempted to model and transfer style from a *multi-speaker database*. They proposed *Mix-StAGE*, a speech-driven approach that trains a model using data from multiple speakers while learning a unique style embedding for each speaker. Their neural architecture uses a *content* and *style* encoder to extract content and style information from pose. A style embedding matrix is used to represent the style associated with each specific speaker in the training set. Their approach presents several limitations. First, behavior expressivity style is only encoded by means of upper-body motion, ignoring the other possible modes of style expression, such as text, speech, and facial expresions. Second, speakers are associated with a unique speaker identity "ID", considered as style labels, which hinder their ability to generalize to new speakers. Later on, Ahuja et al. [Ahuja et al.(2022)] presented a few-shot style transfer strategy based on neural domain adaptation to transfer style with only a few examples, considering the shift in cross-modal grounding between the source speaker and the target style. However, this approach still requires to have at least 2 minutes of the style to be transferred. Fares et. al [Fares et al.(2022), Fares et al.(2023)] proposed an approach to synthesize upper body gestures of a source speaker in the style of any target speaker. The authors do not consider faces in their model. Their approach can be applied to speakers whose style behaviors have been learned or not during the training phase. Overall, the recent models proposed to capture behavior expressivity style have several limitations: they do not exploit multimodal data [Ye et al.(2022), Neff et al.(2008), Alexanderson et al.(2020), Karras et al.(2017), Cudeiro et al.(2019), Ginosar et al.(2019a), Ginosar et al.(2019b), Fares et al.(2022)]; their generative models are trained on single speaker data; style is associated with speaker "IDs", which limits their ability to generalize to new speakers without additional training [Cudeiro et al.(2019), Karras et al.(2017), Ginosar et al.(2019a)]; they require additional training to model unseen target speaker style [Ahuja et al.(2022)]. In addition, they lack a methodology to evaluate properly the behavior expressivity style transfer. In particular, during the style transfer from a target style to a source content, the resulting behavior should ideally preserve the gesture related to the source content (e.g., idiomatic gestures) while modifying its expression accordingly to the target style. Practically, it is a common pitfall in style transfer to observe leakage between the source content and the target style, i.e. partially preserving the source style or transferring the target content.

3 Our contributions

To address the limitations, we introduce *TranSTYLer* a transformer-based model for the generating facial and body gestures of a source speaker in the style of a target speaker, while preserving the intended meaning of the source gestures. Our contributions are:

1. To the best of our knowledge, *TranSTYLer* is the first behavior expressivity style transfer approach that jointly synthesizes *2D upper-body gestures* and *2D facial landmarks* of *source speakers*, in the style of any *target speakers*, and without requiring additional training, making our approach zero-shot.

2. We propose a novel methodology for assessing *behavior expressivity style transfer* for generating communicative behaviors for virtual agents. Our methodology measures *content preservation* and *style transfer* and gives insights about potential leakages between style and content information.
3. We have extended the *PATS corpus*, by including *2D Facial Landmarks* and *Dialog Tags*.

4 *TranSTYler*

We present ***TranSTYler***, a novel approach for modelling *behavior expressivity style* in virtual agents.***TranSTYler*** is a multimodal style transfer approach for generating 2D facial and pose synthesis corresponding to the *content* of a *source speaker* and in the *style* of a *target speaker*. During inference, an embedding style vector can be directly inferred from any target speaker's multimodal data - text, speech, poses, 2D facial landmarks - by simple projection into ***TranSTYler***'s embedding style space (similar to the one used in [Jia et al.(2018)] and [Fares et al.(2023)]). Our approach allows for style transfer from any unseen speakers, without requiring further training or fine-tuning of our trained model. This means that our method is not restricted to the styles of the speakers in the training dataset. ***TranSTYler*** is trained on PATS corpus [Ahuja et al.(2020)] which we extended to include additional facial and text features.

4.1 Problem Positioning

The goal is to learn to generate facial and upper-body gestures based on the source speaker's content information and conditioned on the style information of the target speaker. A transformer-based generator is used to generate facial and body gestures from content and style information. 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, it is discarded, and the model can generate different versions of facial and body gestures when fed with different style vectors. Gesture styles for the same input speech can be directly controlled by switching the value of style vector or by calculating it from a target speaker's multimodal data fed as input to the style encoder. Our approach is based on the following hypotheses:

- Our primary hypothesis is that behavior involves the modulation of communicative gestures associated with content, through the specific gestures associated with an individual. We propose to disentangle this information and encode it in a differentiated manner.
- *Behavior expressivity style* is encoded across *text semantics*, *dialog tags*, *speech*, *face* and *pose* and varies little or not over time. The reason we consider *dialog tags* is to capture further semantic information. Moreover, studies on communicative gestures have shown the link between the meaning carried by dialog acts and the one carried by gestures ([Calbris(2011), Cienki(2005)]).
- *Speech content* is encoded across the verbal and nonverbal modalities. That is, the meaning of what is being said is conveyed by the text and by the nonverbal communicative behaviors.

To implement theses assumptions, we propose an architecture for encoding and disentangling the source speaker's *content* and the target speaker's *style* information from multiple modalities. Style and content information are entangled in each utterance produced by a speaker. On one side, a content encoder $E_{content}$ is used to encode a content matrix from the source's text and speech signal; on the other hand, a style encoder E_{style} is used to encode a style vector from the text, acoustic features, dialog tags, facial and body gestures data of the target. A fader loss ([Lample et al.(2017)]) is introduced to effectively disentangle content and style encodings. The network processes source and target input data at the segment level, where each segment $\boldsymbol{S}$ consists of 64 frames. For each segment $\boldsymbol{S}$, the network takes as input:

1. The source speaker's audio and text semantics represented by the Mel spectrogram (X_{speech}^{source}) and Bert embeddings (X_{text}^{source}).
2. The target speaker's facial gestures, body gestures, audio, text and dialog tags represented by 2D facial landmarks (X_{face}^{target}), 2D poses (X_{body}^{target}), Mel spectrogram (X_{speech}^{target}), Bert embeddings (X_{text}^{target}), and dialog tags (X_{tags}^{target}).

For each $\boldsymbol{S}$, the output of the network is the generation of behaviors that correspond to the content of the source speaker with the style of the target speaker, namely:

1. Facial gestures ($\widehat{Z}_{face}$) represented by 2D facial landmarks.
2. The corresponding upper-body gestures represented by 2D poses ($\widehat{Z}_{body}$).

4.2 Neural Formulation

The network has an embedding size d_{model} equals to 64.

Content Encoder. $E_{content}$ encodes the source speaker's speech content information from the variables X_{text}^{source} and X_{speech}^{source} corresponding to each **S**. X_{text}^{source} is represented by BERT embeddings of dimension 768. X_{speech}^{source} is encoded using $E_{speech}^{content}$, a pre-trained *Mel Spectrogram Transformer (AST) base384* model ([Gong et al.(2021)]), and then concatenated with X_{text}^{source}. 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 speech, which can be written as follows:

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

where: sa(.) denotes self-attention.

Style Encoder. We consider that *behavior expressivity style* is encoded in a speaker's multimodal behavior. As illustrated in Figure 1, E_{style} encodes the behavior expressivity style information from the target speaker's variables X_{speech}^{target}, X_{text}^{target}, X_{pose}^{target}, X_{face}^{target}, and X_{tags}^{target}; which are encoded by E_{speech}^{style}, E_{text}^{style}, E_{pose}^{style}, E_{face}^{style}, and E_{tags}^{style} respectively. E_{pose}^{style} and E_{face}^{style} are composed of N_{lay} equals to 3 layers of LSTMs with a hidden-size equal to d_{model}. E_{tags}^{style} is a *One Hot Encoder* that considers the 38 dialog tags as categorical features. The input features are encoded using a one-hot encoding scheme. The output is a sparse array containing binary values representing the presence or the absence of each tag in the segment **S**. X_{speech}^{target} is encoded by E_{speech}^{style}, which is the pre-trained *AST*. The output vector is concatenated with X_{text} and a self 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 other encoded modalities. 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:

$$\begin{aligned} h_{style} = \big[&sa\left(\left[X_{text}^{target}, E_{speech}^{style}(X_{speech}^{target})\right]\right), \\ &E_{pose}^{style}(X_{pose}^{target}), E_{face}^{style}(X_{face}^{target}), E_{tags}^{style}(X_{tags}^{target})\big] \end{aligned} \tag{2}$$

where: sa(.) denotes self-attention.

Generator. For decoding $\widehat{Z}_{pose}$, and $\widehat{Z}_{face}$, the sequence $h_{content}$ 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 two *Transformer Decoders*. Each *Transformer Decoder* is composed of N_{dec} = *1* decoding layer, with N_h = 2 attention heads, and an embedding size equal to d_{model}. The resulting output vectors are sequences of 2D facial landmarks and 2D-poses which corresponds to:

$$\begin{aligned} \widehat{Z}_{pose} &= G_{pose}(h_{content}, h_{style}) \\ \widehat{Z}_{face} &= G_{face}(h_{content}, h_{style}) \end{aligned} \tag{3}$$

where G_{face} and G_{pose} are the transformer decoders corresponding to *face* and *pose* modalities. The generator loss can therefore be written as:

$$\begin{aligned} \mathcal{L}_{rec}^{gen}(E_{content}, E_{style}, G) = &\mathbb{E}_{\widehat{Z}_{pose}}||Z_{pose} - G_{pose}(h_{content}, h_{style})||_2 \\ &+\mathbb{E}_{\widehat{Z}_{face}}||Z_{face} - G_{face}(h_{content}, h_{style})||_2 \end{aligned} \tag{4}$$

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 h_{style} and $h_{content}$. The latent space of $h_{content}$ is constrained to be independent of h_{style} embeddings. Concretely, it means that the distribution over $h_{content}$ of the latent representations should not contain the style information. We formulate this discriminator Dis as a regression on the conditional variable h_{style}. Dis learns to predict h_{style} from $h_{content}$, 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:

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

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

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

The style prediction error is preliminary normalized within 0 and 1 range. The total G loss can therefore be written as follows:

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

where λ is the adversarial weight that starts off at 0 and is linearly incremented by 0.01 after each training step. The discriminator Dis and the generator G are then optimized alternatively as described in [Lample et al.(2017)]. All ***TranSTYler*** hyperparameters were chosen empirically and are listed in the implementation details section of the appendix.

5 Experimental Evaluations

5.1 Material and Model setups

PATS 2.0 Corpus. The *PATS Corpus* [Ahuja et al.(2020), Ginosar et al.(2019a)] originally includes *2D upper-body joints keypoints*, aligned with the given speech, *Mel spectrogram* and *Bert embeddings*, of 25 speakers, categorized as follows: 15 talk show hosts, 5 lecturers, 3 YouTubers, and 2 televangelists. Each speaker has his/her own communicative style, and lexical and gesture diversity. It has 251 hours of data, with $84,000$ intervals and a mean duration equal to 10.7 seconds per interval. The standard deviation is 13.5 seconds per interval. An interval corresponds to an utterance consisting of 64 timesteps. We have extended PATS to include *2D facial landmarks*,and *dialog tags*, More specifically, we extracted 70 2D facial landmarks for all PATS speakers using OpenPose [Cao et al.(2017)] and aligned with PATS's features. Dialog acts correspond to the communicative functions expressed by the spoken text [Bunt et al.(2010)]. We used the tool "DialogTag" [bha([n. d.])] to extract 38 dialog tags from PATS utterances. We refer the readers to the supplementary materials for the complete list of dialog tags.

TranSTYler Training and Testing. We trained our network using ***PATS 2.0***. Although fingers are included in PATS, we have chosen not to model finger data in our work. The quality of the extracted fingers data is very noisy and lacks accuracy. Instead, we focus on modeling and predicting the 2D joints of the upper body and arms, using 11 joints to represent these areas. We also model 15 facial landmarks, which are illustrated in our Appendix. We use less keypoints than those originally extracted to have less input parameters and fasten the training phase. We took out some keypoints from the face contour and 2 keypoints on each eyebrow, and we used only 2 keypoints for the eyelids. In total, we model 11 body and arm joints, and 15 facial landmarks. An utterance is associated to one or more dialog acts. We consider all the 38 different tags that are listed in our appendix. Our testing comprises two conditions: *Seen Speaker* and *Unseen Speaker*. The *Seen Speaker* condition evaluates how accurately our model can perform style transfer when presented with target speakers seen during training. In contrast, the *Unseen Speaker* condition evaluates our model's ability to perform zero-shot style transfer when presented with target speakers that were not seen during training. We carefully selected both seen and unseen speakers from PATS to cover a range of stylistic behaviors in terms of lexical diversity and spatial extent which is the amplitude of body movements. The PATS database already defines the train, validation, and test sets for each speaker. We train our model on 16 PATS speakers. To test the *Seen Speaker* condition, we use the test sets of the 16 PATS speakers as our test set. For the *Unseen Speaker* condition, we select 6 other speakers and use their test sets for our experiments. Each training batch has BS = 24 pairs of word embeddings, Mel spectrogram, dialog acts, and their corresponding sequence of (x, y) joints of the skeleton of the upper-body pose and 2D facial landmarks. We use Adam optimizer with $\beta_1 = 0.95, \beta_2 = 0.999$, and a *Cyclical Learning Rate* (CLR) scheduler to render the learning balanced. The initial learning rate Lr_{init} of the CLR is equal to $1e-7$, the end learning rate Lr_e is equal to 0.1, and the step size St_{size} is equal to 196. We train the network for N_{it} equals to $78,400$ iterations. All features values are normalized so that the dataset mean and standard deviation are 0 and 0.5, respectively. All hyperparameters used for training are summarized in our appendix.

5.2 Objective Evaluation

We objectively measure the performance of ***TranSTYler*** in terms of two key aspects: *style transfer accuracy* and *content preservation*.Moreover, to measure the degree of similarity of the generated facial and body gestures with the source and target styles, we computed the distance between our model's predictions and each of the source and target styles. Additionally, we assess and compare the unique dynamic movement patterns of the source, target, and prediction by measuring their velocity, acceleration, and jerk. This allows us to quantify and analyze the specific characteristics of their movement dynamics.

Metrics. We have followed the recommendations put forth by Fu et al. [Fu et al.(2018)] in order to evaluate the characteristics of style transfer in our study. We employed their proposed evaluation metrics, *Transfer Strength Accuracy* and *Content Preservation*, to assess the performance of ***TranSTYler***. These metrics measure quantitatively the

accuracy of style transfer and the extent to which content is preserved during the process.

Transfer Strength Accuracy. *Transfer Strength* is a metric that assesses the degree with which the style is transferred. As proposed by Fu et al. [Fu et al.(2018)], this metric is implemented using a classifier C. We consider that *behavior expressivity style* is defined as follows:

$$\textit{Behavior expressivity style} = \begin{cases} \textit{Source (positive) output} \leq 0.5 \\ \textit{Target (negative) output} > 0.5 \end{cases} \tag{9}$$

Transfer Strength Accuracy is defined as follows:

$$\textit{Transfer Strength Accuracy} = \frac{N_{right}}{N_{total}} \times 100 \tag{10}$$

where N_{right} is the number of correct cases which are transferred from target to source style, and N_{total} is the number of test set data. We developed C as a neural network consisting of three LSTM layers and a dense output layer, with the complete architecture shown in the appendix. The network's hyperparameters were chosen empirically and are also listed in the appendix. To train C, we used the train sets of the speakers included in the train sets of both the $Seen$ and $Unseen$ conditions, as defined in the *PATS Corpus*. Specifically, we trained C using a batch size of 256 and $Adam$ optimizer, and a *Binary Cross Entropy* loss over $15,000$ epochs. After training, C achieved an accuracy of 96%.

Content Preservation. *Content Preservation* is a metric that reflects the preservation of source content, that is, in this work, the meaning conveyed by the nonverbal communicative behaviors, in predictions. It is defined as the cosine distance between predictions $\widehat{Z}_{gestures}$ and initial *source* gestures (X_{source}), as follows:

$$\textit{Cosine Distance} = 1 - \frac{X_{source}^{\intercal} \cdot \widehat{Z}_{gestures}}{\|X_{source}\| \|\widehat{Z}_{gestures}\|} \tag{11}$$

Minkowski distance. We also measure the *Minkowski distance* between the upper-body gestures and facial expressions produced by our model, and the ones of the *source* and *target* speakers. We additionally experimented with alternative distance metrics, including *cityblock*, *Chi2 distance*, *Euclidean distance*, and *cosine distance*. However, we found that all these metrics yielded identical results, leading us to retain only the Minkowski distance. More specifically, for both conditions, *Seen* and *Unseen*, we define two sets of distances: (1) ***Dist.(****TranSTYler*, *Source***)** which is the average distance between *TranSTYler*'s predictions and the source's 2D facial landmarks and body joint; and (2) ***Dist.(****TranSTYler*, *Target***)** which is the average distance between *TranSTYler*'s predictions and the target data.

Velocity, Acceleration, Jerk. We evaluate the *velocity*, *acceleration*, and *jerk* of the source, target, and prediction to quantify and compare their distinct dynamic movement patterns. This analysis enables us to determine whether the prediction aligns more closely with the behavior expressivity style of the source or of the target. Velocity provides insights into the overall speed and rhythm of the movement, while acceleration measures the rate of change in motion velocity. Jerk indicates the smoothness of motion transitions over time. By examining these metrics, we can gain valuable information about the characteristics of the movement dynamics and utilize it to assess the degree to which the predicted animation captures the behavior expressivity style of the source or target.

5.3 Human Perceptual Studies

Following previous research [Ahuja et al.(2022), Ahuja et al.(2020)], we contribute to the definition of a comprehensive methodology to assess behavior expressivity style transfer. We focus on differentiating between behaviors associated with the linguistic content of speech (i.e communicative gestures), and the unique style exhibited by a speaker. The desired outcome is to preserve the gestures form associated with the source content while adjusting their expressivity to match the target style. The proposed methodology is defined as follows:

- To assess *behavioral expressivity style transfer*, we evaluate the resemblance of our model's predictions to the target style.
- To assess *content preservation*, we study the coherence of gestures by assessing their coordination with speech content and the synchronization with speech rhythm.

We conduct three studies and compare the perception of stimuli generated by our model and by the two baselines Mix-Stage and DiffGAN.

Study 1. The first study aims to assess the behavior expressivity style produced by our model w.r.t the behavior expressivity style of the *seen* or *unseen* target speakers. We additionally compare our model to *Mix-Stage*[Ahuja et al.(2020)],

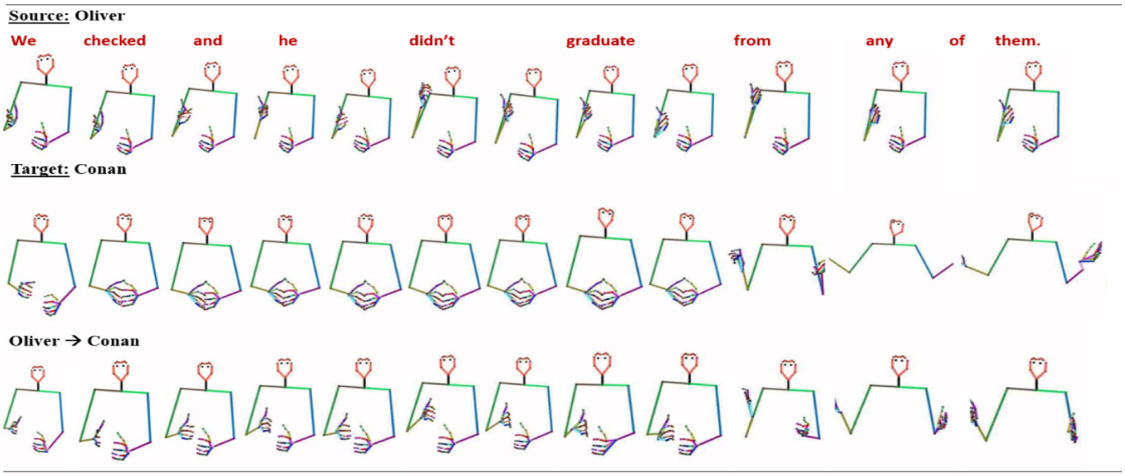


Figure 2: Behavioral expressivity style transfer from target speaker Oliver to source speaker Conan. Fingers are not generated by our model but extracted from OpenPose. They are displayed for sake of visualisation.

which we consider our first baseline. We present 75 stimuli of 2D stick animation (like the 2D sticks in Fig.2) to evaluate our model's predictions with *seen* target styles (*condition 1*, 30 stimuli), our model's predictions with *unseen* target styles (*condition 2*, 30 stimuli), and the baseline *Mix-StAGE* (*condition 3*, 15 stimuli). Each stimulus consists of a triplet of 2D animations composed of: (1) a 2D animation of the source speaker, (2) a 2D animation of the target speaker, and (3) a 2D animation of *TranSTYler*'s prediction after performing the behavioral style transfer. Participants rate on a 5-point Likert scale the *overall resemblance*, *resemblance of the left and right arms gesturing*, *body orientation*, *head orientation*, *gesture amplitude*, *gesture frequency*, and *gesture velocity* of the target style animation with respect to the source style animation and our predictions' animation. The rating scale ranges from 1 (reference is very similar to A) to 5 (reference is very similar to B).

Study 2. We conduct a second study to investigate the coherence of the generated facial and body gestures. Previous research has focused on evaluating the appropriateness of generated gestures [Kucherenko et al.(2023)]. In this work, we place greater emphasis on evaluating the coherence of gestures by assessing their coordination with speech content and synchronization with speech rhythm. By doing so, we aim to subjectively evaluate the extent to which content is preserved after performing behavioral style transfer. We evaluate the coherence of facial and body gestures in relation to speech content and rhythm. We present 90 stimuli of 2D stick animations, comprising 30 stimuli of *TranSTYler*'s stylized predictions (*condition 1*), 30 stimuli of *TranSTYler*'s predictions where we change the original audio with the audio from other speakers (*condition 2*), and 30 stimuli of the source style ground truth (*condition 3*). *Condition 2* is included as an error and control condition. On a 5-point Likert scale, participants rate the *synchronization* of gestures with speech rhythm, the *overall coherence* of behavior, the *coordination* of the agent's gestures with speech content, and the *human-likeness* of the animations.

Study 3. A third study is conducted to compare the similarity of our model's predictions to the target style, as well as to those generated by our second baseline, *DiffGAN* [Ahuja et al.(2022)]. We present 15 stimuli, each comprising a triplet of 2D animations corresponding to the same source-target behavioral style transfer. The first animation is generated by *TranSTYler* (*condition 1*). The second animation represents the reference, and it is *target speakers' ground truth*. The third animation is generated by *DiffGAN* (*condition 2*). For each stimulus, we ask participants to identify which video between *condition 1* and *condition 2* has the same behavior expressivity style as the reference video based on the arm gesture's extent, frequency, timing, and position of the body in relation to speech. We recruited 150 participants through the online crowd-sourcing website Prolific for our evaluation studies. Participants were selected based on their fluency in English. Attention checks were included at the beginning and middle of each study to filter out inattentive participants. Prior to each study, participants received training to introduce them to the 2D facial landmarks and upper-body skeleton and to familiarize them with 2D stick animations.

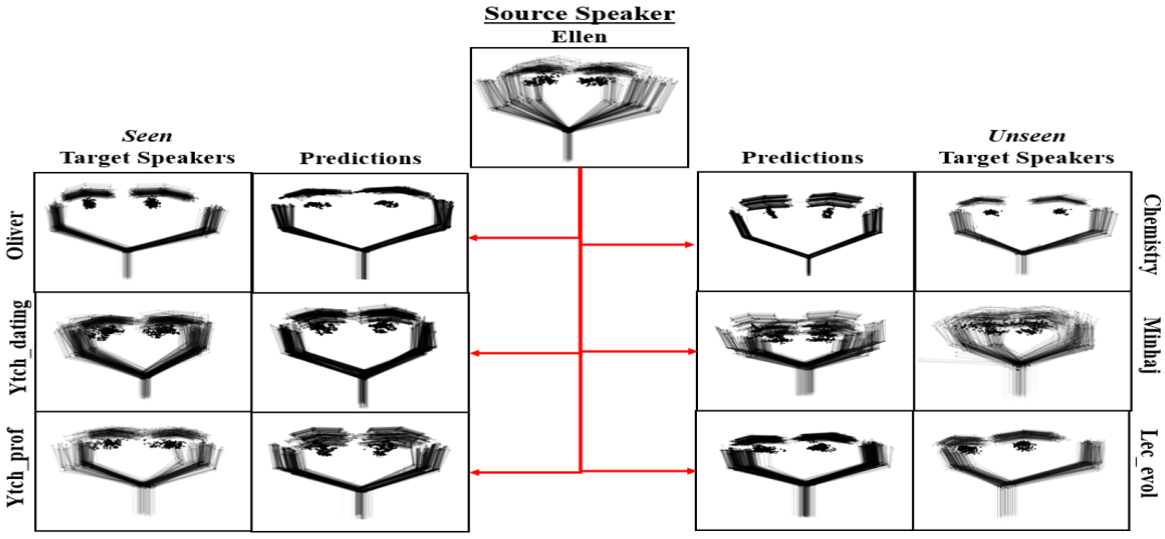


Figure 3: Visualization of facial motion of source, *seen* or *unseen* target and *TranSTYLer*'s predictions.

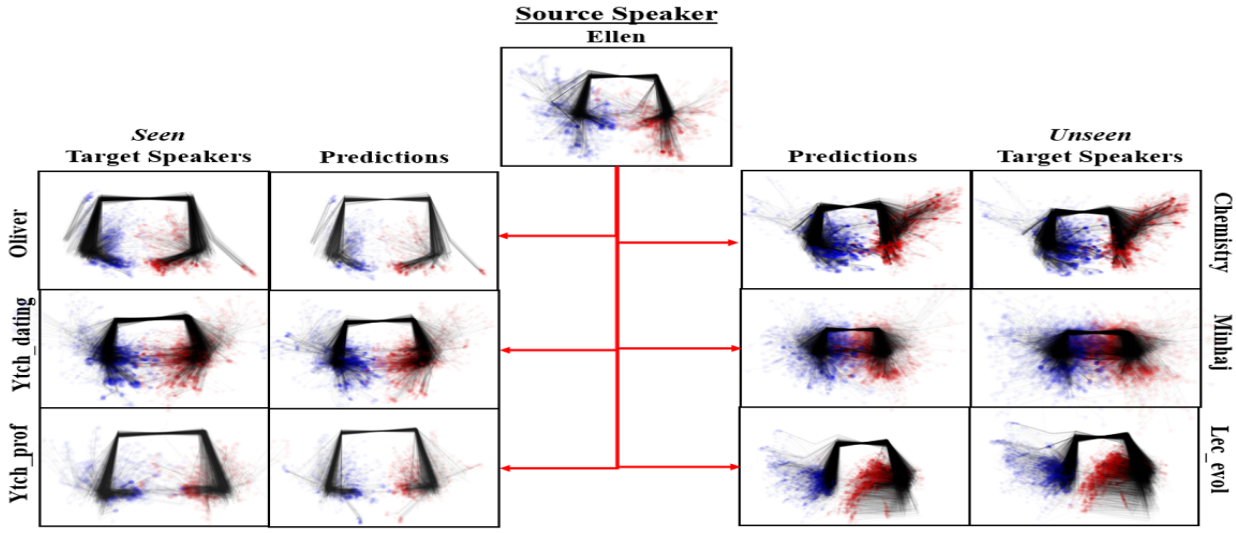


Figure 4: Visualization of upper-body motion of source, *seen* or *unseen* target and *TranSTYLer*'s predictions.

Condition	Transfer Strength Accuracy (%)	Content Preservation (%)	Dist. w.r.t. Source	Dist. w.r.t. Target
Seen	93.282	95.842	83.189	75.882
Unseen	85.195	90.723	80.284	73.934

Table 1: Objective evaluation results: transfer strength accuracy, content preservation, and minkowski distances for Seen and Unseen conditions.

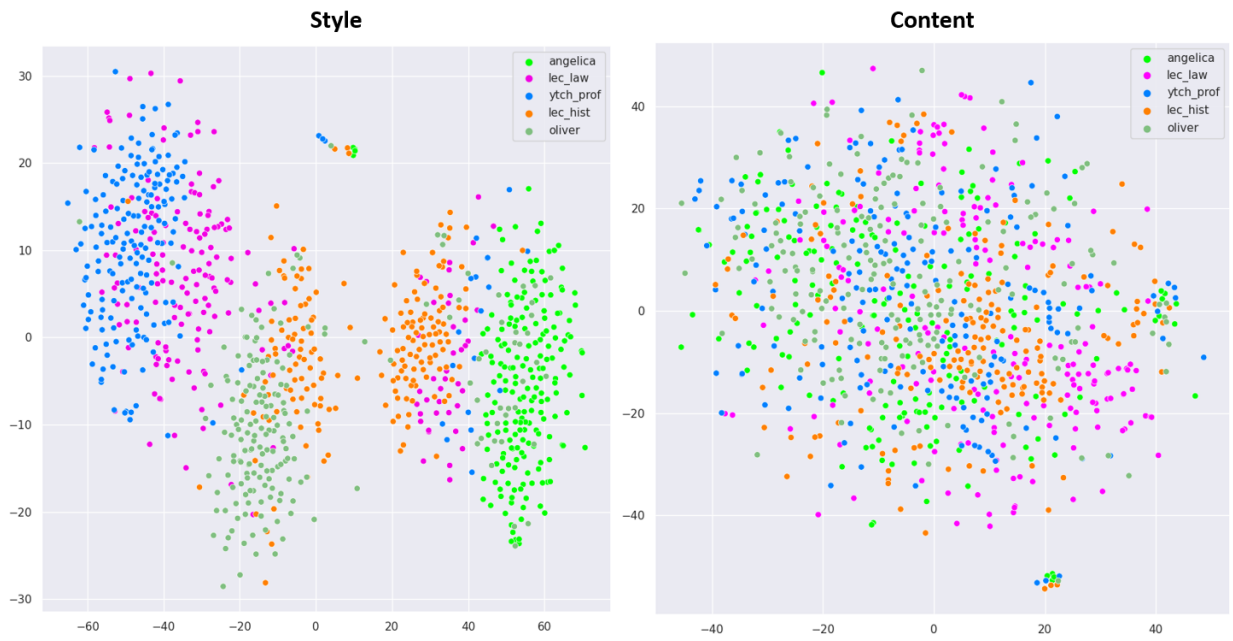
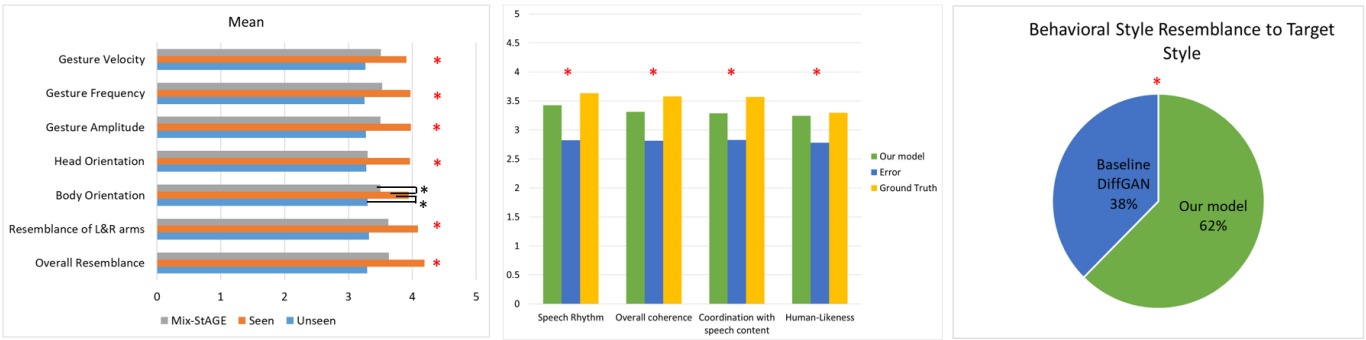


Figure 5: **Left**: T-SNE visualization of the style embeddings on the test sets of 5 speakers. **Right:** T-SNE visualization of the content embeddings on the same test sets.



(a) (b) (c)

Figure 6: Results of perceptual human study 1 (a), study 2 (b), and study 3 (c). Significant differences between pairs of all conditions for the same factor are marked with (∗). If there are significant results between only two pairs of conditions for the same factor, (∗) is used.

6 Results and Discussion

6.1 Objective Evaluation

Objective evaluation results are presented in Table 1 for both *Seen* and *Unseen* conditions. In the *Seen* condition, ***TranSTYler*** achieves a style transfer strength accuracy of 93.282%, indicating a high level of accuracy in transferring the style from the target speakers to the source speakers. For the *Unseen* condition, the accuracy is still high at 85.195%, although slightly lower than the accuracy for the *Seen* condition. This was expected since ***TranSTYler*** had not seen the target speakers during training. Nonetheless, the model demonstrated the ability to generalize the style to new, unseen speakers. In both the *Seen* and *Unseen* conditions, our model is able to preserve a high percentage of the source speakers' content, with 95.842% and 90.723% content preservation, respectively. The distance between our model's predictions and the source speakers' gestures - ***Dist.(TranSTYler,* Source)** - is higher than the one between our model's predictions and the target speakers' gestures - ***Dist.(TranSTYler,* Target)**. These results confirm that the *behavior expressivity style* is successfully transferred from *target* to *source* speakers for both conditions. These

results are further illustrated in Figures 3 and 4 that illustrate the facial and body motion of a same source speaker (Ellen) as well as different target speakers that were either seen or unseen during training, alongside *TranSTYler*'s predictions after performing style transfer. Figure 2 illustrates the motion of a source speaker Oliver, a target (seen) speaker Conan, and that of *TranSTYler*'s predictions at the frame level. The source speaker Oliver gestures mainly with his right hand while Conan makes ample arm movements as shown toward the end of his sentence. The predicted animation displays the communicative gestures of Oliver (similar vertical movement of the arm) and the amplitude extent of Conan toward the end of the sentence. To explore the relationships and patterns between the content and style vectors generated by our model, a 2D T-SNE analysis was conducted. This analysis projects the vectors onto a two-dimensional space, where their proximity indicated similarity. The TSNE plots in Figure 5 showcase the content embeddings ($h_{content}$) and style embeddings (h_{style}) after disentangling the style-content information. By examining the distribution of features in the content and style space, it was observed that content and style were effectively separated. The style space exhibited clustering of features belonging to the same speaker, suggesting discernible patterns. However, in the content space, features from all speakers were mixed together without clear patterns or clusters. These results demonstrate the success of our disentangling approach in effectively separating style-content information. We additionally computed the velocity, acceleration, and jerk of *TranSTYler*'s predictions, source style, and target style for source-target style transfers where the target is either *Seen* or *Unseen*. The results for style transfers performed on four source-target pairs, with two *Seen* targets (Angelica and Lec_hist) and two *Unseen* targets (Almaram and Minhaj) indicate that, for both *Seen* and *Unseen* targets, *TranSTYler*'s velocity is closer to the target than to the source. Regarding the acceleration metric, we observed similar results for all style transfers except for the transfer from source ***Bee*** to the unseen target ***Minhaj***, where the predictions' acceleration is closer to the source style. However, for the same ***Bee - Minhaj*** style transfer, our predictions' jerkness is closer to the target than the source style. For the style transfer ***Lec_law - angelica***, ***TranSTYler*** produces a velocity that is close to the target style and far from the source style, an acceleration that is in between the source and target style, and a jerkness closer to the source style. Overall, these findings show that *TranSTYler* effectively transfers the behavior expressivity style from the target to the source speakers, as evidenced by the high style transfer accuracy, content preservation, and the observed patterns in velocity, acceleration, and jerk metrics.

6.2 Subjective Evaluation

Figure 6 (a) shows the mean scores obtained for all factors for all conditions (*Mix-StAGE*, *Seen* and *Unseen*); the higher the mean score, the closer the condition is w.r.t the target style. For all factors, our model obtained the mean scores higher than those of the baseline. For all factors, *Mix-StAGE* received lower scores than the *Seen* condition and higher scores than the *Unseen* condition. This may be due to the fact that speakers are visible during training in the *Mix-StAGE* condition, whereas *TranSTYler* is unseen in the *Unseen* condition. We conducted post-hoc paired t-tests for each factor between the three conditions and found significant differences ($p < 0.007$) between *Mix-StAGE* and *Seen*, and *Unseen* and *Seen* for all factors. We found significant results for *Mix-StAGE* and *Unseen* for all factors except 'body orientation'. Our prediction model in both *Seen* and *Unseen* conditions outperforms the baseline for all factors. The *Seen* condition also surpasses the *Unseen* one. Our prediction model in both *Seen* and *Unseen* conditions outperforms the baseline for all factors. The *Seen* condition also surpasses the *Unseen* one. The goal of Study 2 is to assess the preservation of the speech content during the style transfer. It evaluates the coherence of gestures by examining their coordination with speech content and synchronization with speech rhythm. Results of **Study 2** are presented in Figure 6(b). We conducted paired t-tests for each factor between the following conditions: *TranSTYler* and *Error*, *TranSTYler* and *Ground Truth*, and *Ground Truth* and *Error*. The results showed significant differences ($p < 0.001$) between each pair of conditions for all factors. *TranSTYler* achieved scores that are significantly ($p < 0.001$) very similar to the ground truth scores. In contrast, the control condition, *error*, obtained a significantly ($p < 0.001$) lower mean score than the scores obtained by our model. Thus the gestures computed by our model maintained adequacy with the speech content as predicted gestures are highly similar to those in the original video. However, we are aware that further study ought to be conducted on measuring more precisely the semantic and pragmatic information conveyed by the predicted behaviors. The third study aimed to compare our model, *TranSTYler*, with a second baseline, *DiffGAN*[Ahuja et al.(2022)], and results are shown in Figure 6 (c). Participants were asked to identify which animation, between condition 1 (*TranSTYler*) and condition 2 (*DiffGAN*), had the most similar behavior expressivity style as the reference video (*target style*) based on the arm gesture's extent, frequency, timing, and position of the body in relation to speech. A post-hoc binomial test was also conducted, and significant results were found for both conditions ($p < 0.001$). Thus, overall, our model generates animations that significantly capture better the behavior expressivity style of the target speaker than does *DiffGAN*.

7 Conclusion

We present *TranSTYLer* for synthesizing body and facial gestures of source speakers in the style of target speakers, without additional training. We propose a novel methodology for evaluating behavior expressivity style transfer, measuring content preservation and style transfer while identifying potential leakages between style and content information. Furthermore, we expand the PATS corpus by including 2D Facial Landmarks and Dialog Tags.

References

[bha([n. d.])] [n. d.]. *GitHub* ([n. d.]). `https://github.com/bhavitvyamalik/DialogTag`

[Ahuja et al.(2022)] 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)*.

[Ahuja et al.(2020)] 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.

[Alexanderson et al.(2020)] 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.

[Alibali et al.(2009)] Martha W Alibali, Julia L Evans, Autumn B Hostetter, Kristin Ryan, and Elina Mainela-Arnold. 2009. Gesture–speech integration in narrative: Are children less redundant than adults? *Gesture* 9, 3 (2009), 290–311.

[Argyle(2013)] Michael Argyle. 2013. *Bodily communication*. Routledge.

[Armstrong et al.(1995)] David F Armstrong, William C Stokoe, and Sherman E Wilcox. 1995. *Gesture and the nature of language*. Cambridge University Press.

[Bunt et al.(2010)] Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, et al. 2010. Towards an ISO standard for dialogue act annotation. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*.

[Calbris(2011)] Geneviève Calbris. 2011. *Elements of Meaning in Gesture*. John Benjamins Publishing Company.

[Cao et al.(2017)] 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.

[Cienki(2005)] Alan Cienki. 2005. Image schemas and gesture. *From perception to meaning: Image schemas in cognitive linguistics* 29 (2005), 421–442.

[Cudeiro et al.(2019)] 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.

[Fares et al.(2022)] Mireille Fares, Michele Grimaldi, Catherine Pelachaud, and Nicolas Obin. 2022. Zero-Shot Style Transfer for Gesture Animation driven by Text and Speech using Adversarial Disentanglement of Multimodal Style Encoding. *arXiv preprint arXiv:2208.01917* (2022).

[Fares et al.(2023)] Mireille Fares, Catherine Pelachaud, and Nicolas Obin. 2023. Zero-Shot Style Transfer for Multimodal Data-Driven Gesture Synthesis. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–4.

[Feyereisen et al.(1991)] Pierre Feyereisen, Jacques-Dominique De Lannoy, et al. 1991. *Gestures and speech: Psychological investigations*. Cambridge University Press.

[Feyereisen and Havard(1999)] Pierre Feyereisen and Isabelle Havard. 1999. Mental imagery and production of hand gestures while speaking in younger and older adults. *Journal of nonverbal behavior* 23, 2 (1999), 153–171.

[Fu et al.(2018)] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[Ghorbani et al.(2022a)] Saeed Ghorbani, Ylva Ferstl, and Marc-André Carbonneau. 2022a. Exemplar-based stylized gesture generation from speech: An entry to the GENEA Challenge 2022. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 778–783.

[Ghorbani et al.(2022b)] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F Troje, and Marc-André Carbonneau. 2022b. Zeroeggs: Zero-shot example-based gesture generation from speech. *arXiv preprint arXiv:2209.07556* (2022).

[Ginosar et al.(2019a)] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019a. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3497–3506.

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

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

[Hostetter and Alibali(2007)] Autumn B Hostetter and Martha W Alibali. 2007. Raise your hand if you're spatial: Relations between verbal and spatial skills and gesture production. *Gesture* 7, 1 (2007), 73–95.

[Hostetter and Potthoff(2012)] Autumn B Hostetter and Andrea L Potthoff. 2012. Effects of personality and social situation on representational gesture production. *Gesture* 12, 1 (2012), 62–83.

[Jia et al.(2018)] 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).

[Karras et al.(2017)] 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.

[Kita(2009)] Sotaro Kita. 2009. Cross-cultural variation of speech-accompanying gesture: A review. *Language and cognitive processes* 24, 2 (2009), 145–167.

[Knapp et al.(2013)] Mark L Knapp, Judith A Hall, and Terrence G Horgan. 2013. *Nonverbal communication in human interaction*. Cengage Learning.

[Kucherenko et al.(2023)] Taras Kucherenko, Pieter Wolfert, Youngwoo Yoon, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. Evaluating gesture-generation in a large-scale open challenge: The GENEA Challenge 2022. *arXiv preprint arXiv:2303.08737* (2023).

[Lample et al.(2017)] 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).

[McCrae and Costa Jr(1997)] Robert R McCrae and Paul T Costa Jr. 1997. Personality trait structure as a human universal. *American psychologist* 52, 5 (1997), 509.

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

[Neff et al.(2008)] 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.

[Ye et al.(2022)] Sheng Ye, Yu-Hui Wen, Yanan Sun, Ying He, Ziyang Zhang, Yaoyuan Wang, Weihua He, and Yong-Jin Liu. 2022. Audio-Driven Stylized Gesture Generation with Flow-Based Model. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*. Springer, 712–728.

[Yoon et al.(2020)] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–16.