Zero-Shot Style Transfer for Multimodal Data-Driven Gesture Synthesis

Mireille Fares, Nicolas Obin, Catherine Pelachaud

Zero-Shot Style Transfer for Multimodal Data-Driven Gesture Synthesis

Mireille Fares[1,2], Catherine Pelachaud[2], and Nicolas Obin[1]
[1] STMS Lab, IRCAM, Sorbonne Université, France
[2] ISIR Lab, Sorbonne Université, France

***Abstract*— We propose a multimodal speech driven approach to generate 2D upper-body gestures for virtual agents, in the communicative style of different speakers, *seen* or *unseen* by our model during training. Upper-body gestures of a *source speaker* are generated based on the *content* of his/her multimodal data - speech acoustics and text semantics. The synthesized *source speaker*'s gestures are conditioned on the multimodal style representation of the *target speaker*. Our approach is zero-shot, and can generalize the style transfer to new *unseen speakers*, without any additional training. An objective evaluation is conducted to validate our approach.**

I. Introduction

Human behavioral style involves the ways in which speakers gesture in various situations, which can carry different social meanings [5]. It is *multimodal* and present in all human communication modalities: text semantics [7], speech prosody [21], [23], and gestures (i.e hand, facial and body postures) [22], [28]. *Verbal and Non-verbal behavior* define speaker's communication style behavior, which is adapted during social interactions to the style of others [13]. *Prosodic features* which are present in human voice are related to the way specific sounds are produced [4], [6]. In this paper, we present the first approach to generate semantically-aware and speech-driven 2D upper-body gestures of a *source speaker* in the style of different target speakers, including those unseen during training. The *source speaker*'s gestures are generated based on the *content* of two input modalities - Mel spectrogram and text semantics. The generated gestures are conditioned on a multimodal representation of a *target speaker* style. Our approach is zero-shot, it allows style transfer of speakers unseen during training without any further training or fine-tuning. We assume that behavioral *style* is multimodal, could be encoded using the three modalities: text, speech and upper-body gestures, and that that *content* information is only found in the *content* of text and speech modalities. We apply these hypothesis by disentangling *style* from *content* while relying on the fader network disentangling approach [20]. We validate our approach by conducting an objective evaluation.

979-8-3503-4544-5/23/$31.00 ©2023 IEEE. 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.

II. Related Work and Our Contributions

Many gesture generative models have been previously proposed. Long-Short Term Memories (LSTMs) [31], [30], Generative Adversarial Networks (GANs) [26], [25], [12], and Convolutional Neural Networks (CNNs) combined with GANs [16] were used to pblackict *speech-driven* head motion, facial, hand and body gestures. Other works driven by both *speech and text semantics* were proposed for upper-facial [11], [9], [10] and hand [19] gesture generation, however they cannot be used for style modelling and control. Synthesizing expressive gestures while controlling their style is recently receiving more attention [3], [17], [8], [14]. However, none of the works consider the *multimodal* aspect of style in their models.The only works that model and transfer speakers style using a *multi-speakers* database have been proposed by Ahuja et al. in [2] and [1]. They propose [2] a speech-driven style-transfer model trained on PATS - a database for studying unique styles of gestures for a large number of speakers. In their proposed architecture, they assume that *style* is only present in speakers gestures, and do not consider its presence in speech and text. Their model is limited to PATS speakers, and is not *zero-shot* since it cannot generalize the style transfer to new *unseen* speakers. They propose another approach[1] based on a few-shot style transfer strategy that uses domain adaptation between *source speaker* and *target speaker* style. Nonetheless, this adaptation needs 2 minutes of training to perform the style transfer. Our contributions can be listed as follows: (1) We propose the first approach for zero-shot multimodal style transfer. At inference, a style vector can be inferblack from multimodal data of any *target speaker*, and is used to condition the synthesis of 2D upper body gestures of a *source speaker*. Our model is trained on PATS dataset, but the style transfer is not limited to PATS speakers. It can generalize to new *unseen* speakers. (2) Behavioral *style*

is encoded using the three modalities: text, speech and 2D upper-body gestures. *Content* is encoded only by speech and text *content* of *source speakers*.

III. Proposed Architecture

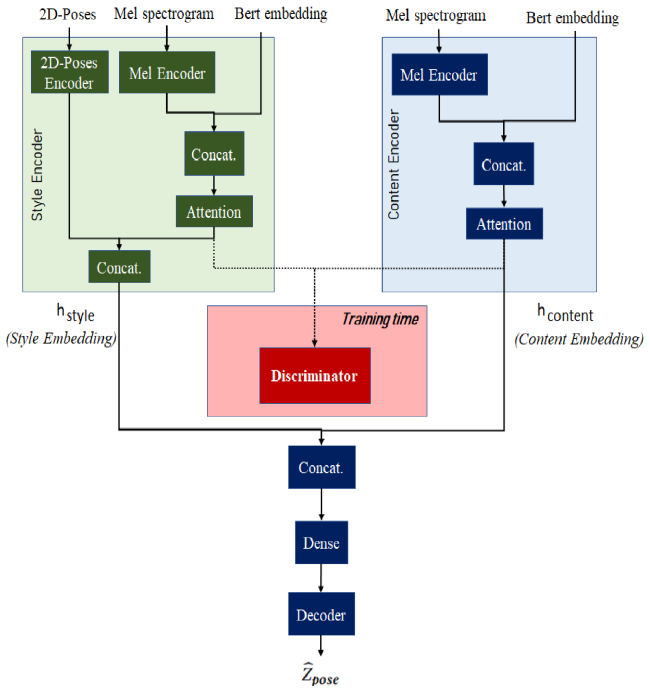


Fig. 1: **ZS-MSTM** model architecture.

We propose **ZS-MSTM** (**Z**ero-**S**hot **M**ultimodal **S**tyle **T**ransfer **M**odel), a transformer-based architecture for synthesizing 2D upper-body gestures of a *source speaker* in the style of a *target speaker*. Gestures are generated based on the content of the *source speaker*'s speech - text semantics (BERT embeddings) and audio (Mel spectrogram) -, and conditioned on the *target speaker*'s multimodal style embedding generated by the *target speaker*'s Mel spectrogram, BERT embeddings, and 2D upper-body gestures, which are represented by a sequence of (X, Y) joints positions. The generated *animations* correspond to the style of target speakers *seen* or *unseen* during training. The network operates on an utterance-level. The input features - text and speech - corresponds to an utterance $\boldsymbol{U}$ with a length of *64* timesteps, as provided by PATS. The model synthesizes a sequence of gestures that correspond to the given $\boldsymbol{U}$ input features. As depicted in Fig.1, the model is composed of: (1) a content encoder, (2) a style encoder, and (3) an adversarial component. During inference, the adversarial component is removed, and the model can synthesize various styles of gestures when fed with different style embeddings. Gesture styles for the same input speech can be controlled by changing the style embedding vector h_{style} or by computing this embedding from a target speaker's multimodal data given as input to the *Style Encoder*.

A. Content Encoder

The content encoder takes as inputs a sequence of BERT embeddings X_{text} and an audio Mel spectrograms Y_{speech} corresponding to each utterance **U**. Y_{speech} is encoded using AST pre-trained model [15], from which we removed the final linear layer with sigmoid activation function, since we do not aim to classify audio. The encoded Mel spectrogram is then concatenated with the BERT embeddings of $\boldsymbol{U}$. Self attention is applied on the resulting vector. The output is $h_{content}$, a representation of the content of the source speaker's speech and text semantics corresponding to $\boldsymbol{U}$.

B. Style Encoder

The style encoder E_{style} takes as input the *target speaker*'s Mel spectrogram Y_{speech}, BERT embedding X_{text}, and a sequence of (X, Y) joints positions corresponding to $\boldsymbol{U}$. The modified *AST* is used to encode Y_{speech}. A self-attention is then applied on the resulting vector. 3 layers of LSTMs with a hidden-size equal to 768 are used to encode the 2D poses. The last hidden layer is then concatenated with the audio representation. The output embedding h_{style} is the *target speaker* style embedding that serves to condition the network with the speaker style.

C. Sequence to sequence gesture synthesis

This network is a generator G, that generates the sequence of poses $\widehat{Z}_{pose}$ from the sequence of $h_{content}$ and h_{style}. To decode the stylized 2D-poses, $h_{content}$ and h_{style} are concatenated, and passed through a dense layer. The resulting vector is then given as input to a transformer decoder [27] with 1 decoding layer and 2 attention heads. The resulting vector $\widehat{Z}$pose is the sequence of stylized 2D-poses.

D. Adversarial Component

An adversarial component in the form of a *fader network* [20] is used to disentangle *style* and *content*. The goal is to constrain the latent space of $h_{content}$ to be independent of h_{style}, such that the distribution over $h_{content}$ doesn't contain style information. We formulate this problem as a regression on the conditional variable h_{style}: the discriminator D learns to pblackict h_{style} from $h_{content}$. While optimizing D, the discriminator loss $\mathcal{L}^{dis}$ must be as low as possible. While optimizing the generator loss, D must not be able to predict correctly h_{style} from $h_{content}$ conducting to a high D

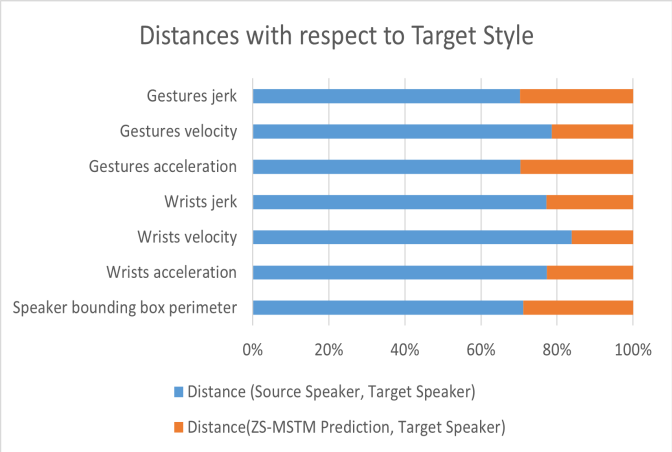
error and thus a low fader loss. D and G are then optimized alternatively as described in [20].

IV. Objective Evaluation

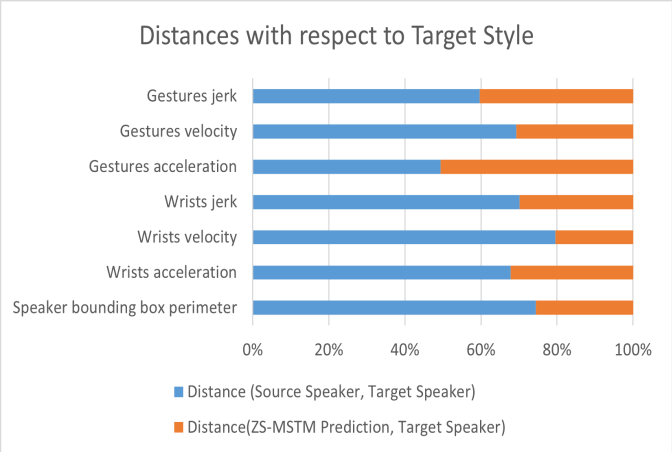
A. Objective Evaluation

We conduct an objective evaluation to assess our model in transferring the style of *seen* and *unseen* speakers. Style corresponds to the quality of behavioral expressivity. Following works on behavior expressivity by [29], [24], we use 4 objective behavior dynamics metrics: acceleration, jerk and velocity and the speaker's Bounding Box (BB) perimeter. To obtain information on the arms movements expressivity, we compute the acceleration, jerk and velocity of the left and right wrists ([29], [18]).

B. Objective Evaluation Results and Discussion



(a) Style transfer from seen speakers



(b) Style transfer from unseen speakers

Fig. 2: Average distance between source and target styles compablack to the one between our model's pblackictions and target styles

Objective evaluation experiments are conducted for evaluating the performance of our model in two conditions: *Seen* and *Unseen* conditions. For *Seen* condition, experiments are conducted on a test set that includes the 16 speakers *seen* by our model during training. For *Unseen* condition, experiments are conducted on another test set with 6 *unseen* speakers. We define two sets of distances: ***Dist.****(Source*, *Target***)** - average distance between the source and target styles -, and ***Dist.****(ZS-MSTM*, *Target***)** - average distance between our model's pblackictions and the target styles. Fig. 2 (a) reports the experimental results on the *Seen* test set. It illustrates the results of ***Dist.****(Source*, *Target***)** in terms of behaviors dynamics and speaker BB perimeter between the target and the source speakers styles. The ***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 perimeter of the pblackiction's BB is closer (dist < 30 %) to the target speaker's BB perimeter than the source. The closeness between pblackictions dynamics behavior metrics values are shown for all speakers in the *Seen* condition. Results for the *Unseen* test set are shown in Fig. 2 (b)For all behavior dynamics metrics, as well as the BB 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 closer to the target speaker than the source one. The BB perimeter of *ZS-MSTM*'s pblackiction's is close to the target speaker's BB perimeter(dist. < 30 %), while there is a larger distance (dist. > 70 %) between the source and target speakers' BB perimeter. Wrists acceleration and jerk values of our model's produced gestures are very close to those of the target speakers for the 6 *unseen* speakers. We additionally found significant results ($p < 0.003$) for all distances in both conditions, *seen* and *unseen*, after conducting a Fisher's LSD Test. We conclude that our model is capable of transferring behavior expressivity style of all *seen and unseen target speakers* to source speakers.

V. Conclusion and Future Work

We presented the first approach for zero-shot multimodal style transfer for 2D upper-body gestures synthesis. We plan to conduct subjective evaluations on virtual agents, and expand our model to consider dialog acts, and cover facial gestures.

REFERENCES

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

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

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

[4] W. Apple, L. A. Streeter, and R. M. Krauss. Effects of pitch and speech rate on personal attributions. *Journal of personality and social psychology*, 37(5):715, 1979.

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

[6] D. B. Buller and R. K. Aune. The effects of speech rate similarity on compliance: Application of communication accommodation theory. *Western Journal of Communication (includes Communication Reports)*, 56(1):37–53, 1992.

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

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

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

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

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

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

[13] H. Giles. *Communication accommodation theory: Negotiating personal relationships and social identities across contexts*. Cambridge University Press, 2016.

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

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

[16] I. Habibie, W. Xu, D. Mehta, L. Liu, H.-P. Seidel, G. Pons-Moll, M. Elgharib, and C. Theobalt. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 101–108, 2021.

[17] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017.

[18] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 97–104, 2019.

[19] T. Kucherenko, P. Jonell, S. van Waveren, G. E. Henter, S. Alexanderson, I. Leite, and H. Kjellström. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the ACM International Conference on Multimodal Interaction*, 2020.

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

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

[22] C. Obermeier, S. D. Kelly, and T. C. Gunter. A speaker's gesture style can affect language comprehension: Erp evidence from gesture-speech integration. *Social cognitive and affective neuroscience*, 10(9):1236–1243, 2015.

[23] N. Obin. *MeLos: Analysis and modelling of speech prosody and speaking style*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2011.

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

[25] M. Rebol, C. Gütl, and K. Pietroszek. Real-time gesture animation generation from speech for virtual human interaction. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–4, 2021.

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

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

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

[29] H. Wallbott. Bodily expression of emotion. *European Journal of Social Psychology*, 28:879–896, 1998.

[30] J. Woo. Development of an interactive human/agent loop using multimodal recurrent neural networks. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 822–826, 2021.

[31] J. Woo, C. Pelachaud, and C. Achard. Creating an interactive human/agent loop using multimodal recurrent neural networks. In *WACAI 2021*, 2021.