Multimodal Expressive Gesturing With Style

PhD Candidate: Mireille Fares
Thesis Director: Catherine Pelachaud
Co-advisor: Nicolas Obin
Keywords: Multimodality, Gesture SynThesis, Style Transfer, Transformer Networks, Visual Prosody, ECAs

Thesis Context

Figure 1 - Greta (Pelachaud [2015])

Embodied Conversational Agents (ECAs) (*Figure 1*) are computer-generated human-like characters that originated from the idea of simulating multimodal human communication where the modalities are those of human communication. They have many of the same properties as humans during conversation, including the capacity to generate and respond to verbal and non-verbal signals (Ball and Breese [2000]). Although it seems simple for humans, ECAs must properly understand the mechanisms that govern *human multimodal behavior* to produce similar behavioral expressivity.

The central theme of this Thesis is to model the relationships existing between *human visual prosody*, *speech prosody*, and *spoken verbal language*, to build ECAs that can produce *coherent* and *human-like* non-verbal behavior. The first objective is to exploit these multimodal mechanisms and develop generative models that can synthesize *expressive visual prosody* for ECAs. The second objective is to model human multimodal *behavioral style* to control and adapt the ECA's *behavioral style* to any *behavioral style* of a *target speaker*, by performing a *style transfer* from the *target speaker* to the ECA's synthesized behavior. Below we dive deeper into the different objectives of this Thesis.

Objective 1. How can we model relationships between visual prosody, speech prosody, language, and the speech context (text semantics) so that we can synthesize human-like multimodal gestures ? It is challenging to answer this question and understand these relationships due to the complex relationships present in human verbal and non-verbal signals that are emitted during speech. Expressive visual prosody is indeed an issue in current generative models. The first objective of this Thesis aims to develop a model that can synthesize human-like expressive visual prosody. The gestures that are produced by this model correspond to the *multimodal data* given as input, particularly *text semantics* and *speech prosody*. The expressivity and human-likeness of the produced gestures are leveraged by effectively exploiting the human multimodal behavior data given as input to the model. The goal is to render the generated gestures human-like, synchronized with the given input speech and aligned with the text content.

Objective 2. Gestures have unique semiotic properties, they are idiosyncratic, and are created locally by speakers during their speech (McNeill et al. [2005]). Each speaker has a different *gesturing* and *behavioral style* when communicating. Speakers discussing same topics, or belonging to the same category of speakers (i.e. TV hosts, lecturers, televangelists, etc.) may have common gestural style, as mutual *behavioral style* can generally be found between these speakers. Modeling *behavioral style* is a key challenge and a complex problem, since it is *multimodal* (Knapp et al. [2013]), and found in *verbal* and *non-verbal* behavior (Campbell-Kibler et al. [2006], Moon et al. [2022], Obin [2011], Obermeier et al. [2015], Wagner et al. [2014]). The second objective of this Thesis is to model *human behavioral style*. The goal is to be able to control the style of the synthesized gestures, and perform *style transfer* amongst different speakers that are *seen* by our model, and *unseen* as well (zero-shot style transfer). For instance, consider two different speakers *Alice* and *Bob*. *Alice* is speaking with her own gesturing style. *Bob* is likely to gesticulate differently than *Alice*. Our goal is to be able to transfer the gestural style from *Alice* to *Bob*, so that *Bob* can follow the same gesturing style as *Alice* while speaking any utterance. Modeling *behavioral style* is necessary to learn a style space based on multimodal speakers data. The challenge is to build this style space, that is independent from speakers' identity (which in most previous works is defined by their "ID"), and is only dependent on speakers' *multimodal data*, which were *seen* by our model during training. Another challenge is to generalize style to new *unseen* speakers, such that the model is able to generalize *behavioral style* to *unseen* speakers, without any further training or fine-tuning, thus allowing us to perform *zero-shot style transfer*.

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. A variety of generative statistical models aimed to predict the multimodal behavior of a virtual agent. HMMs (Hofer and Shimodaira [2007]), Recurrent Neural Networks (RNN) (Wang et al. [2021], Haag and Shimodaira [2016]), and Dynamic Bayesian Networks (DBN) (Mariooryad and Busso [2012], Sadoughi and Busso [2019]) have been used to generate head motion from speech; Generative Adversarial Networks (GAN) have been proposed to produce facial gestures from speech (Karras et al. [2017], Vougioukas et al. [2019]). The main limitation of most of these works is that they exploit as input only one modality, namely *speech*, and neglect to render them *semantically-aware*. In addition to that, most of them focus on modeling only one type of gesturing, without considering the correlation existing between several types of gesturing. For instance, facial expressions and head movements are highly correlated to prosody - more specifically, the fundamental frequency f0 (Yehia et al. [2002]) - and therefore are highly correlated to each other. Modeling them together is crucial to produce a natural behavior in ECAs, which was not previously done.

Beyond the realistic generation of human non-verbal behavior, *style modeling* and *control* in gesture is receiving more attention in order to propose more expressive behaviors that could possibly be adapted to a specific audience. A large number of other generative models were proposed in the past few years for synthesizing gestures in the style of specific speakers. Some of these works generate full body gesture animation driven by *text*, and 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. [2019]) are *speech-driven*, and they were proposed for generating gesticulation in a certain style. For some approaches, the style of the synthesized gestures is changed by exerting direct control over the synthesized gestures' velocity and levels (Alexanderson et al. [2020]). For others (Cudeiro et al. [2019], Karras et al. [2017], Ginosar et al. [2019]), they produce the gestures in the style of a single speaker by training their generative models on a single speaker's data, and synthesizing the gestures corresponding to the speaker's specific audio. The main limitation of these works is that they have focused on generating gestures (facial expression, head movement, gestures in particular) that are aligned with either *speech* or *text*. They did not exploit multimodal data for their *gesture synthesis*, nor for modeling the *behavioral style* of speakers. Moreover, their generative models are trained on one *single speaker data*. The only attempts to model and transfer the *style* from a *multi-speakers* database (Ahuja et al. [2020b] and Ahuja et al. [2022]) are only *speech*-driven. Their approach does not exploit verbal information for synthesizing the gestures. Moreover, their models are conditioned on the behavioral style that is only found in the speakers gesturing. However, as previously discussed, behavioral style is also found in other modalities such as speech prosody and linguistics. In their approach, behavioral style is associated with each unique speaker identity, which makes the distinction unclear between each speaker's specific behavioral style, and the behavioral style that is common between specific speakers, for instance speakers that are TV show hosts, or journalists. In addition, their approach is limited to the behavioral styles of the speakers that are in the database that was used for training. Their approach cannot generalize the behavioral style to new speakers without the need of additional training and fine-tuning of their model (Ahuja et al. [2022]). They perform neural domain adaptation between a source speaker style and a specific target speaker style.

Contributions

For the purpose of addressing the different previously mentioned limitations and technical challenges, this Thesis proposes two corpora and several models to synthesize stylized co-expressive facial and body gestures accompanying speech, synchronized with it, semantically-aware, and aligned with the speech content. The contributions of this Thesis are listed as follows:

Corpora. The first contribution of this Thesis is the development of the *TEDx database* for the purpose of using it in our different studies related to multimodal *gesture synthesis*. *TEDx database* includes multimodal features extracted from 1760 TEDx talks that were collected. These features consist of *facial gestures*, *acoustic*, and *text semantics* features. Another contribution is the extension of *PATS* database (Ahuja et al. [2020b]) for the purpose of using it in our different studies related to human *behavioral style*. *PATS* was first proposed by Ahuja et al. [2020b], and initially included *body features*, *acoustic* and *text* features of different speakers having different *behavioral styles*. We extended it to include *facial gestures features* (*2D facial landmarks*), and *dialog acts.*

Semantically-aware and speech-driven upper facial gestures. The second contribution of this Thesis is the development of an approach for upper-facial synthesis based on *speech prosody* and *text semantics*. Eyebrow motion is represented by facial Action Units, and their intensities as described by the FACS (HAGER [2002]). We exploit both *speech* and *text* modalities to generate co-expressive human-like eyebrow motion. We conduct several objective and subjective evaluations to validate our approach. To assess objectively the quality of the generated gestures, we used the following metrics: *Root Mean Squared Error (RMSE)*, *Pearson Correlation Coefficient (PCC)*, *Activity Hit Ratio (AHR)* and *Non-Activity Hit Ratio (NAHR)*. Subjective evaluations aimed to assess the *appropriateness*, *coherence*, *naturalness*, *synchronization* and *alignment* of the synthesized gestures with the given input modalities data. We show that the usage of both modalities leverages the quality of the results.

Head motion synthesis. Besides synthesizing co-expressive *eyebrow motion*, we extend our model to generate *head motion* as well. We additionally model the correlation between *head motion* and *upper facial gestures* at the output of our model to synthesize *coherent* and *natural* behavior of the agent. Additional objective and subjective evaluations are conducted to validate the performance of our approach.

Zero-shot style transfer - for body pose synthesis. The fourth contribution of this Thesis is the development of ***ZS-MSTM***, a model that allows zero-shot multimodal style transfer for 2D body pose synthesis. This model produces stylized upper-body gestures, driven by the *content* of a *source* speaker's speech - text semantics 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 can be seen or unseen during training. This model allows us to directly infer an embedding style vector from multimodal data (text semantics, dialog tags, speech and pose) of any speaker, by simple projection into the embedding style space. The style transfer performed by our model allows the transfer of style from any *unseen* speakers, without further training or fine-turning, rendering our approach *zero-shot*. Objective and subjective evaluations are conducted to evaluate *ZS-MSTM* and validate it.

Zero-shot style transfer - for facial and body pose synthesis. The last contribution of this Thesis is an extension of ***ZS-MSTM*** that includes the synthesis of *facial landmarks*, leading to a model that can synthesize *body pose animation* with their corresponding and *facial landmarks*. This model is the first to synthesize 2D body poses and facial landmarks altogether, while learning the *behavioral style*. Similar to the other models, objective and subjective evaluations are conducted to evaluate the expressivity of the generated facial and body visual prosody.

Publications and Submissions

Fares, M., Pelachaud, C., & Obin, N. (2022, August). Transformer Network for Semantically-Aware and Speech-Driven Upper-Face Generation. In *EUSIPCO*.

Fares, M. (2020, October). Towards multimodal human-like characteristics and expressive visual prosody in virtual agents. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (pp. 743-747).

Fares, M., Pelachaud, C.., & Obin, N. (2020, June). Multimodal modeling of expressiveness for human-machine interaction. In *Workshop sur les Affects, Compagnons Artificiels et Interactions*.

Fares, M., Pelachaud, C., & Obin, N. (2021, October). Multimodal-Based Upper Facial Gestures SynThesis for Engaging Virtual Agents. In *WACAI*.

(Accepted, to appear) **Fares, M., Pelachaud, C. & Obin, N. Zero-Shot Style Transfer for Multimodal Data-Driven Gesture SynThesis. In SIVA'23.**

(Preprint) - **Fares, M., Grimaldi, M., Pelachaud, C., & Obin, N. (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*.**

(Paper under review at HCII'23) - **Fares, M., Pelachaud, C. & Obin, N., I-Brow: Hierarchical and Multi-Modal Transformer Model for Eyebrows Animation SynThesis**

(Submission in progress at Frontiers in AI) - **Fares, M., Pelachaud, C. & Obin, N., Zero-Shot Style Transfer for Body Gesture Synthesis**