Behavior Generation Model for Socially Interactive Agents

Mireille Fares[1], *Catherine Pelachaud* [2], *Nicolas Obin* [3]

[1] ISIR, STMS, Sorbonne University

[2] CNRS, ISIR, Sorbonne University

[3] STMS, Sorbonne University

fares@isir.upmc.fr, catherine.pelachaud@isir.upmc.fr, nicolas.obin@ircam.fr

We are interested in modeling non-verbal communicative behaviors for Socially Interactive Agents SIAs. We have proposed models to capture gestures characteristics, their time of occurrence, shape and movement. One of our focus has been on metaphoric gestures where our approach was based on the concept of Image Schemas to capture the underlying physical properties of an idea to be translated into gesture characteristics. Lately, we have developed deep learning approaches to learn the relation between text semantics, speech acoustic features and gesture, and to simulate the communication with different behavior styles. These models learn well the temporal relationship between speech acoustics and gesture timing, but less regarding gesture shape. In this abstract we present our different endeavors and future directions.

SIAs are virtual entities, often with a human-like appearance, that are autonomous and able to communicate verbally and nonverbally with human users [Lugrin et al., 21]. Over the years, several approaches have been proposed to compute which multimodal behaviors the agent should display to convey its intentions. The first models were rule-based relying on theoretical models [Cassell et al., 01] and statistically-based [Bergman and Kopp, 09]. Lately, data-driven approaches have been applied. These models capture the temporal relationship between speech prosody and behaviors, such as head movement, eyebrows shape and arm gestures. They are trained on large databases of either single speaker or multiple ones [Kucherenko et al., 20; Fares et al., 22].

In our group, we have developed several models to allow SIAs to communicate through its multimodal behaviors. One of our approach focused on metaphoric gestures [Cienki, 05; Lakoff and Johnson, 80]. We were particularly interested in computing the hand shape and arm movement that correspond to the underlying physical properties of metaphors [Ravenet et al., 18]. We relied on embodied cognition theory for which gesture production arises from a common representation in verbal and nonverbal modalities. Our model is a simplification of theoretical models [Lakoff and Johnson, 80] in that Image Schemas ISs are directly extracted from speech and not arising from cognitive processes (see Figure 1). Then, gesture properties are defined by the extracted ISs. For example, the IS OBJECT will be converted into the conduit shape while the IS UP will set the wrist position high. The next step is to combine the gestures characteristics to render the final arm animation of virtual agents. We use Calbris' Ideational Units concept to instantiate all the undefined gesture components [Calbris, 11]. This model allows us to compute gestures shape and motion aligned with the semantics of speech. Such an approach gives animations that carry some form of meaning but lack of naturalness. Indeed, it instantiates gestures only where ISs are extracted.

On the other hand, our latest models rely on data-driven approaches and render more natural animation. Our architecture relies on transformer networks and attention mechanisms that are performant for transforming sequences (here speech) into other sequences (multimodal behaviors) [Fares et al., 22]. To encode semantic information, we use BERT semantic encoding, while we use F0 to encode speech prosody. The output of this architecture is facial expression and head motion linked to speech. We trained our model on a large database made of TEDx video. Such an approach allows generating facial gestures synchronized with speech.

So far, we were interested in generating nonverbal behaviors linked to speech. However, we all gesture differently. We have our own behavior style that may come from a large range of factors such as our personality traits, emotions, role, environment, etc. Our aim is to generate the nonverbal behaviors of a virtual agent with the style of a specific person. We consider that style is embedded not only in the nonverbal behaviors, but also in the spoken language, namely acoustic features and text. Our model learns the behavior style of different persons from their multimodal behaviors (body pose and facial expression), their prosody and text. Then it applies a similar sequence-to-sequence model as in [Fares et al., 22] to generate nonverbal behaviors but conditioned by a given behavior style (see Figure 2, [Fares et al., 23]).

To validate our models, we have defined objective measures that compute temporal and spatial errors of body animation compared to the ground truth (extracted animation from original videos). We have also conducted perceptive evaluation studies to find out if the generated animations are perceived as being synchronized with the speech and if they look natural.

These last two models generate animations that preserve the rhythm of nonverbal behaviors aligned with speech, producing natural and expressive animations. However, they do not reproduce the gesture shape with high fidelity. Using BERT encoding for text does not offer a rich enough semantic representation to be translated into gesture shape. In the next future, our aim is to encapsulate semantic information, such as ISs, in our deep learning architecture. As a first step it requires proposing a common representation of speech and gesture and using such a representation in our architecture.

Index Terms: Socially Interactive Agent, communicative behavior style, non-verbal behavior generation

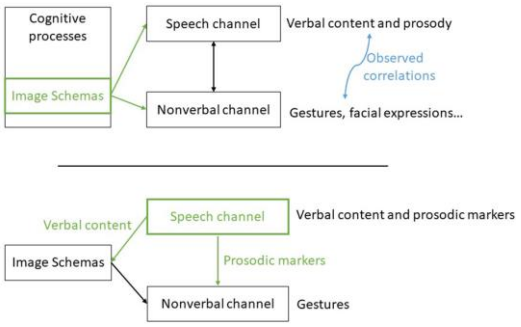

Figure 1: *top figure: simplified theoretical model of verbal and nonverbal production [Kendon, 80]; bottom figure: Ravenet's behavior generation model [Ravenet et al., 18]*

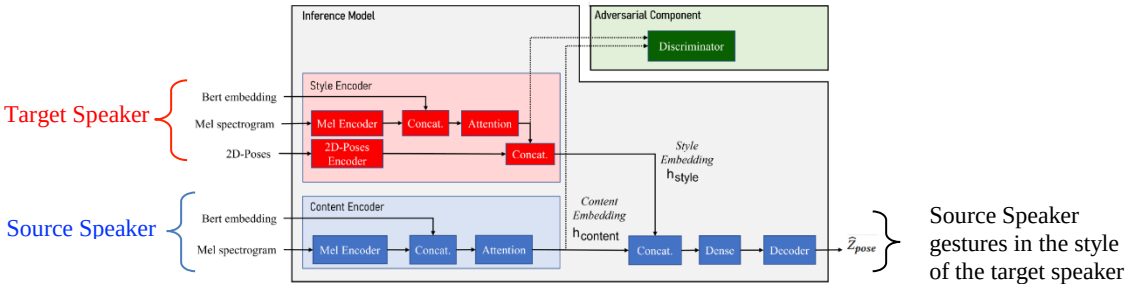

Figure 2: *Overall architecture of style transfer model [Fares et al., 2023]: the blue box encodes what is being said using BERT word embedding and Mel spectrogram. The red box encodes the behavior style to be transferred. It outputs the multimodal behavior linked to the content of speech of the Source Speaker with the style of the Target Speaker.*

References

Bergmann, K., & Kopp, S. 2009. Gnetic–using bayesian decision networks for iconic gesture generation. In *International Workshop on Intelligent Virtual Agents*, pp. 76–89. Springer.

Calbris, G. 2011. *Elements of Meaning in Gesture*, John Benjamins, 1-398.

Cassell, J., Vilhjálmsson, H., & Bickmore. T. 2001. BEAT: the Behavior Expression Animation Toolkit. In *Computer Graphics Proceedings, Annual Conference Series*. ACM SIGGRAPH.

Cienki, A. 2005. Image schemas and gesture. In Hampe, B. & Grady, J.E. (Eds), *From perception to meaning: Image schemas in cognitive linguistics*, 29, Mouton de Gruyter, 421-442.

Fares, M., Pelachaud, C., & Obin, N. 2022. Transformer Network for Semantically-Aware and Speech-Driven Upper-Face Generation. In *30th European Signal Processing Conference* (EUSIPCO) (pp. 593-597). IEEE.

Fares, M., Pelachaud, C., & Obin, N. 2023. Zero-shot style transfer for gesture animation driven by text and speech using adversarial disentanglement of multimodal style encoding. *Frontiers in Artificial Intelligence*. 6:1142997.

Kendon, A. 1980. "Gesture and speech: two aspects of the process of utterance, in Key, M.R. (Ed), *Nonverbal Communication and Language*, The Hague: Mouton, 207–227.

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.*

Lakoff, G., & Johnson, M. 1980. Conceptual metaphor in everyday language. *Journal of Philosophy*, 77(8), 453-486.

Lugrin, B., Pelachaud, C., & Traum, D. 2020-22. The handbook on socially interactive agents: 20 years of research on embodied conversational agents, intelligent virtual agents, and social robotics, ACM.

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