Multimodal-Based Upper Facial Gestures Synthesis for Engaging Virtual Agents

Mireille Fares, Catherine I Pelachaud, Nicolas Obin

Multimodal-Based Upper Facial Gestures Synthesis for Engaging Virtual Agents

Mireille Fares
ISIR and IRCAM
Sorbonne Université
Paris, France
fares@isir.upmc.com

Catherine Pelachaud
CNRS-ISIR
Sorbonne Université
Paris, France
catherine.pelachaud@upmc.fr

Nicolas Obin
CNRS-IRCAM
Sorbonne Université
Paris, France
nicolas.obin@ircam.fr

ABSTRACT

Myriad of applications involve the interaction of humans with machines, such as reception agents, home assistants, chatbots or autonomous vehicles' agents. Humans can control the virtual agents by the mean of various modalities including sound, vision, and touch. In this paper, we discuss about designing engaging virtual agents with expressive gestures and prosody. We also propose an architecture that generates upper facial movements based on two modalities: speech and text. This paper is part of a work that aims to review the mechanisms that govern multimodal interaction, such as the agent's expressiveness and the adaptation of its behavior, to help remove technological barriers and develop a conversational agent capable of adapting naturally and coherently to its interlocutor.

KEYWORDS

multimodality, speech, gestures, prosody, intelligent embodied conversational agents, neural networks

1 INTRODUCTION

Human-Human interaction inherently involves the communication through multiple channels. We employ several modalities, both sequentially and in parallel, to communicate in our daily life. The multimodal channels adopted in human communication are verbal and non-verbal [17]. Both verbal and non-verbal modalities are essential to send and perceive new information. A key problem in the design of virtual assistants is how to maintain user's engagement [3, 4] during the interaction so that the interaction lasts long and stays fluent. The present paper is part of a thesis work that aims to better understand and model the mechanisms that govern multimodal interaction (voice and gesture) between a human and a machine. We aim to develop engaging embodied conversational agent (ECA) with expressive gestures and prosody. As a first step of our work, we present in this paper a review of how to design engaging virtual agents with expressive gestures and prosody, and we propose an architecture that generates upper facial gestures based on prosody.

2 ENGAGING EMBODIED CONVERSATIONAL AGENTS

Engagement is "the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake" as defined by Sidner et al [20]. The design of Embodied Conversational Agents (ECA) capable of rendering the users engaged in the interaction, is essential and critical for Human Agent Interaction applications [9]. The behavior of the agent should adapt to the multimodal behavior of the interlocutor [8, 21]. Short-term engagement is needed for performing a specific task during the interaction, whereas long-term engagement is essential for longer periods of interactions [2]. The development of an engaging ECA must take into account the socio-emotional behavior of users which are expressed by means of verbal or non-verbal signals [9]. A socio emotional behavior includes user's social attitudes as well as their emotions. As Scherer defines it, a social attitude is "an affective style that spontaneously develops or is strategically employed in the interaction with a person" [19]. An engaging agent detects the engagement level of its interlocutor and maintains it by displaying an appropriate socio-emotional behavior during the interaction. Some studies prefer to detect the user's level of engagement by analyzing his/her signals (such as the location of his/her face), instead of his/her socio emotional behavior [5]. Other studies prefer to focus on acoustic features such as prosody, voice quality, and spectral features..

3 RELATED WORK

Facial Gesture synthesizing systems for virtual agents has received a lot of attention in the past years. For instance, in [7], they generate expressive facial movements that are synchronized with the acoustic features of the input utterances. In [22], they generate lower facial movements using a deep learning approach: they use a sliding window predictor that learns mappings from phonemes to mouth movements. The authors of [26], present a facial gesture generation system. In their approach, virtual speakers takes input text and generate appropriate speech and facial movements. Hofer [13] proposes a speech driven head motion sequence prediction system that uses Hidden Markov Models (HMM). Moreover, in [12], the authors propose a technique for speech driven head motion synthesis that uses deep neural networks with stacked bottleneck features, along with an LSTM network. Mariooryad and Busso [16] propose a facial animation system based on speech to synthesize head and eyebrows motion using Dynamic Bayesian Networks.

4 UPPER-FACIAL GESTURES SYNTHESIS

This section presents part of a work that aims to generate the virtual agent's upper facial gestures, based on speech. The agent's visual prosody expressiveness is modelled using Recurrent Neural Networks (RNN) that are recently commonly used for modelling voice and gestures [18, 24, 25]. RNNs are used to model gestural variability at the Inter-Pausal Unit (IPU) level, as well as the word level. An Inter-Pausal Unit assigns boundaries before and after pauses that are longer than 0.2 seconds. There are two segmentation units used in this work: word-level, and IPU-level.

We have built an end-to-end neural network model architecture that consists of several encoders to encode the input features, and a decoder to generate the sequence of action units that are related to eyebrow and eyelid movement. These action units are described in Facial Action Coding System (FACS) [11]: a system that defines facial movements based on 44 Action Units (AUs). Our model is trained on a series of TEDx talks [23]: TED (Technology, Entertainment, Design) are conferences where experts share their major research and ideas from myriad of disciplines with their audience.

The different modalities and features that we consider in our model are: speech audio, text, and action units.

- Action Units features: AU01, AU02, AU04, AU05, AU06 and AU07 which represent upper facial gestures. They were extracted using OpenFace [1]. We applied a median filter with a window size of 7 to each action unit intensity, to eliminate openface noises.
- Audio features: the audio feature that we are considering in our model is the prosodic feature the fundamental frequency F0. F0 was extracted using SWIPE estimator [6]. IrcamAlign [15] was used to generate the alignment of speech signals into phones and diphones which are pairs of phonetic sounds that are adjacent to each other in an utterance. It also provides a confidence level for each phone.
- Text features: a speech text is a sequence of words that is aligned with its corresponding action units and audio features. Each word is encoded as a BERT embedding [10].

The problem we want to solve consists in mapping a sequence of text, as well as a sequence of fundamental frequencies to a sequence of action units. Since this problem looks very similar to the machine translation problem, we decided to implement a sequence to sequence model to solve our problem. The network architecture is composed of two different parts: the encoder and the decoder. The encoder is illustrated in Figure 1. It takes as input the input words and the input sequence of F0s that corresponds to each word one by one. These words and Fundamental Frequencies correspond to one interpausal unit. This network includes several Bi-directional LSTM layers, as well as 1D convolutional layers. A cross attention mechanism is added to the encoder, to render the network able to learn when to focus on one modality more than the other.

The output of the encoder is then transmitted to the decoder part of the network. The decoder uses a dot product attention mechanism. This part of the network includes LSTM layers, as well as 1D convolutional layers. The output of the decoder is a sequence of action units that corresponds to the same sequence of words and F0s given as inputs to the network.

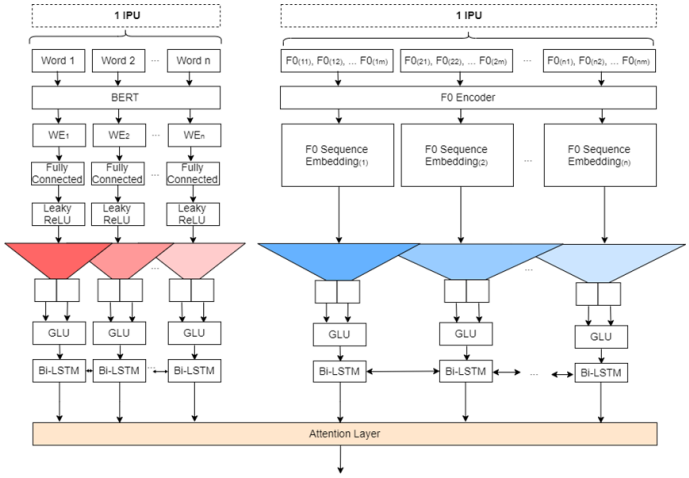

Figure 1: Encoder Network Architecture.

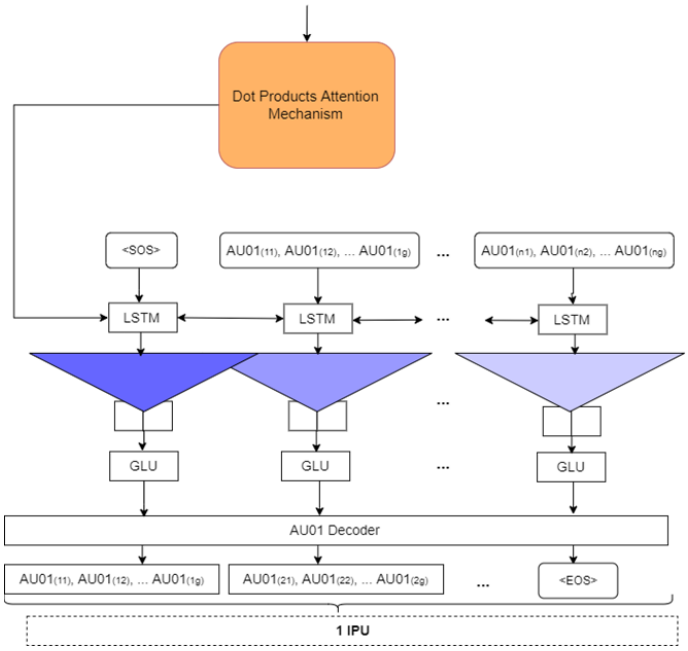

Figure 2: Decoder Network Architecture.

5 FUTURE WORK

We have described our latest work which is the development of an architecture that models gestural and prosodic expressiveness: it predicts sequences of upper facial movements based on the fundamental frequency. For future work, we plan to conduct different types of experiments on our architecture: ablation experiments as well as perceptive experiments. The ablation experiments consist of removing some parts of the network, and testing to see if the quality of predictions is not deteriorated. The perceptive experiments are conducted by generating simulations on the Embodied Conversational Agent GRETA [14]: these simulations will include the upper facial gestures generated by our model, as well as the raw gestures extracted from TEDx talks. Participants will be asked to evaluate the coherence, expressiveness, and naturalness levels of these simulations.

ACKNOWLEDGMENTS

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.

REFERENCES

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.

[2] Timothy Bickmore and Toni Giorgino. 2006. Health dialog systems for patients and consumers. *Journal of biomedical informatics* 39, 5 (October 2006), 556—571. https://doi.org/10.1016/j.jbi.2005.12.004

[3] Timothy Bickmore, Daniel Schulman, and Langxuan Yin. 2010. Maintaining engagement in long-term interventions with relational agents. *Applied Artificial Intelligence* 24, 6 (2010), 648–666.

[4] Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and Maintaining Long-Term Human-Computer Relationships. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (June 2005), 293–327. https://doi.org/10.1145/1067860.1067867

[5] Dan Bohus and Eric Horvitz. 2014. Managing Human-Robot Engagement with Forecasts and... Um... Hesitations. In *Proceedings of the 16th International Conference on Multimodal Interaction* (Istanbul, Turkey) *(ICMI '14)*. Association for Computing Machinery, New York, NY, USA, 2–9. https://doi.org/10.1145/2663204.2663241

[6] Arturo Camacho and John G Harris. 2008. A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America* 124, 3 (2008), 1638–1652.

[7] Yong Cao, Wen C Tien, Petros Faloutsos, and Frédéric Pighin. 2005. Expressive speech-driven facial animation. *ACM Transactions on Graphics (TOG)* 24, 4 (2005), 1283–1302.

[8] Ginevra Castellano, Maurizio Mancini, Christopher Peters, and Peter W McOwan. 2011. Expressive copying behavior for social agents: A perceptual analysis. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 42, 3 (2011), 776–783.

[9] Chloé Clavel, Angelo Cafaro, Sabrina Campano, and Catherine Pelachaud. 2016. *Fostering User Engagement in Face-to-Face Human-Agent Interactions: A Survey*. Springer International Publishing.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[11] Rosenberg Ekman. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

[12] Kathrin Haag and Hiroshi Shimodaira. 2016. Bidirectional LSTM networks employing stacked bottleneck features for expressive speech-driven head motion synthesis. In *Int. Conference on Intelligent Virtual Agents*. Springer, 198–207.

[13] Gregor Hofer and Hiroshi Shimodaira. 2007. Automatic head motion prediction from speech data. (2007).

[14] Isir. [n.d.]. isir/greta. https://github.com/isir/greta

[15] Pierre Lanchantin, Andrew C Morris, Xavier Rodet, and Christophe Veaux. 2008. Automatic Phoneme Segmentation with Relaxed Textual Constraints.. In *LREC*.

[16] Soroosh Mariooryad and Carlos Busso. 2012. Generating human-like behaviors using joint, speech-driven models for conversational agents. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 8 (2012), 2329–2340.

[17] Sigrid Norris. 2004. *Analyzing multimodal interaction: A methodological framework*. Routledge.

[18] Carl Robinson, Nicolas Obin, and Axel Roebel. 2019. Sequence-to-sequence Modelling of F0 for Speech Emotion Conversion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6830–6834.

[19] Klaus R Scherer. 2005. What are emotions? And how can they be measured? *Social science information* 44, 4 (2005), 695–729.

[20] CL Sidner. 2004. Where to look: A study of human-robot interaction. In *Proc. International Conference on Intelligent User Interfaces (ACM IUI 2004)*. 78–84.

[21] Candace L. Sidner, Cory D. Kidd, Christopher Lee, and Neal Lesh. 2004. Where to Look: A Study of Human-Robot Engagement. In *Proceedings of the 9th International Conference on Intelligent User Interfaces* (Funchal, Madeira, Portugal) *(IUI '04)*. Association for Computing Machinery, New York, NY, USA, 78–84. https://doi.org/10.1145/964442.964458

[22] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–11.

[23] TEDxTalks. [n.d.]. TEDx Talks. https://www.youtube.com/user/TEDxTalks

[24] Xin Wang, Shinji Takaki, and Junichi Yamagishi. 2017. An RNN-Based Quantized F0 Model with Multi-Tier Feedback Links for Text-to-Speech Synthesis.. In *INTERSPEECH*. 1059–1063.

[25] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017* (2018).

[26] Goranka Zoric, Karlo Smid, and Igor S Pandzic. [n.d.]. Automated Gesturing for Embodied Animated Agent: Speech-driven and Text-driven Approaches. *Journal of Multimedia* 1, 1 ([n. d.]).