Generation of Multimodal Behaviors

Michele Grimaldi
ISIR, Sorbonne University
Paris, France
michele.grimaldi@isir.upmc.fr

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

KEYWORDS

virtual agent, FML, image schema, ideational unit, multimodal behavior

1 INTRODUCTION

One of the main challenges when developing Embodied Conversational Agents ECAs is to give them the ability to autonomously produce meaningful and coordinated verbal and nonverbal behaviors. Those behaviors concern the movement of some body parts, simultaneously or not, that usually suits the context of the speech, accentuate its effect or give it another interpretation.
There are different models that can compute multimodal behaviors for virtual agent. Our goal is to merge some of them to have a large range of behaviors to create expressive ECAs.

2 GRETA

Greta [8] is a real-time three dimensional embodied conversational agent with a 3D model of a woman compliant with MPEG-4 animation standard. It is able to communicate using a rich palette of verbal and nonverbal behaviors. Greta can talk and simultaneously show facial expressions, gestures, gaze, and head movements.

Two standard XML languages FML [4] and BML [5] allow the user to define its communicative intentions and behaviors based on the standard SAIBA architecture. Originally, Greta standard treatment instantiates the communicative intentions defined in an FML file into multimodal behaviors. The instantiation is based on a lexicon that contains pairs of the type (intention, multimodal behaviors). The instantiated behaviors are then combined and synchronized with the speech; they are sent to the Behavior Realizer that produces the final animation of the virtual agent.
That process is at the basis of all Greta's behaviors but is limited in terms of behaviors complexity. Our current goal is, given communicative intentions specified in an FML file, to generate multimodal behaviors relying one three different computational models and then to give a priority to the behaviors that were computed.
We present now the two other modules that compute automatically multimodal behaviors, which we have added.

3 GESTURES GENERATION FROM IMAGE SCHEMA

When communicating one shows a variety of communicating gestures that can depict the shape of an object, indicate a point in space or convey an abstract idea.
Ideational Units are units of meaning that give rhythm to the discourse of a person and during which gestures show similar properties. A gesture can have invariant properties that are critical for its meaning. For instance, a gesture representing an ascension would probably have an upward movement or an upward direction but the hand shape may not be of particular relevance (for conveying the ascension meaning)[10]. Within an Ideational Unit, successive gestures need to show significant changes to be distinguished.
Image Schemas are the mental representation of what is being conveyed [1].
We have developed a module called Meaning Miner [9] based on the concepts of Images Schemas[1] and Ideational Unit as the intermediate language between the verbal and nonverbal channels.

Meaning Miner reads the FML file marked with prosodic and Ideational Unit [10]. Then it finds the corresponding image schemas and builds the corresponding gesture. The Image Schemas Extraction module has the task of identifying the Image Schemas from the surface text of the agent's speech and to align them properly with the spoken utterance (for future gesture alignment). After obtaining a list of aligned Image Schemas for a sequence of spoken text, the gesture modeler module builds the corresponding gestures. The first step is to retrieve the gesture invariants to build the final gestures. Gesture invariants are the features that need not to be altered to properly express a given meaning [10]. Gestures are described by several features, namely the hand shape, orientation, movement and position in gesture space [6] (Bressem, 2013).
For each Image Schema we search which features are needed to express its meaning and how it is expressed using a dictionary that maps each Image Schema into its corresponding invariants.

Meaning Miner has to co-articulate gestures within an Ideational Unit, that structures speech, by computing either a hold or an intermediate relaxed pose between successive gestures (instead of returning to a rest pose), transfers properties of the main gesture onto the variant properties of the other gestures of the same Ideational Unit ensures that a meaning expressed through an invariant is carried on the same hand throughout an Ideational Unit and finally dynamically raises the speed and amplitude of repeated gestures.

4 NVBG

To extend the range of behaviors that Greta's agent can assume, the NVBG (Non Verbal Behavior Generator) [7] module has been integrated in the Greta platform. NVBG analyzes the syntactic and semantic structure of the agent text as well as the affective state and computes the corresponding nonverbal behaviors.
Greta launches NVBG and its charniak parser [2]. It sends the text to be said to the NVBG module as a vrExpress Message. In NVBG gestures are distinguished by:

- animation: it concerns the whole body movement excepted the head
- head: head movement such as nod or shake

In NVBG, an animation is a macro-type that conveys communicative acts such as: contemplate, negation, contrast, response, request, listening and so on. If NVBG finds an animation or head movement it will send back the encrypted response of the treatment.

The animation tag, the macro-type and the name of the gesture that NVBG outputs are not recognized in the Greta platform. The animations in NVBG are motion capture data. For that reason the NVBG's response is treated to suits the XML elements and gesture names understandable by Greta. For that purpose there are two mapping files that are used to map the name of NVBG's gesture and types to Greta's gesture and types.

Each animation outputed by NVBG is an XML line that has some attributes. Firstly the mapping files allow us to change the NVBG type and name with the corresponding gesture in Greta. Then the attributes that are not useful for Greta are removed or modified. NVBG defines only the start of a gesture at a certain time marker without defining when it should finish (as it is defined by the corresponding motion capture animation file). Thus, when translating NVBG to Greta, an end attribute is added for each animation line. The full treated response is a series of XML lines with tags linked to FML entries (eg deictic, iconic, performative) that are understandable by Greta.

Now the Behavior Planner can understand the gestures computed from NVBG. It then needs to translate them into BML signals. The whole process can be summarized as a transformation of mocap animation into BML signals. Finally the Behavior Realizer receives all the signals and computes the animation via the MPEG-4 player.

5 INTEGRATION

The integration that has been done allows Greta to use Meaning Miner and NVBG models independently and to combine the results of their treatments with the standard treatment that is done by default. For that purpose the two software were included not as external modules but permanently inside Greta as their treatment needs to be used constantly. The Greta architecture has been updated to allow it to access NVBG and Meaning Miner treatment. The two models are added as external modules. The FML module reads an FML file and interacts in series with the three behavior generation modules. It stores the gestures that are computed by each of them. Then it sends them to the Behavior Planner that orders and synchronized them.

The Behavior Planner is re-designed to contain the NVBG and Meaning Miner Models. The rest of the process, translation of the gesture into Signals and then signals into video-audio animation, is not affected. The Behavior Planner now contains all the treatments. Even if one of them does not find gestures this does not affect the rest of the process.

6 CONCLUSION AND FUTURE WORKS

The redesigned architecture integrates several modules to compute automatically multimodal behaviors from a semantic analysis as well as the analysis of the underlying mental representation of what

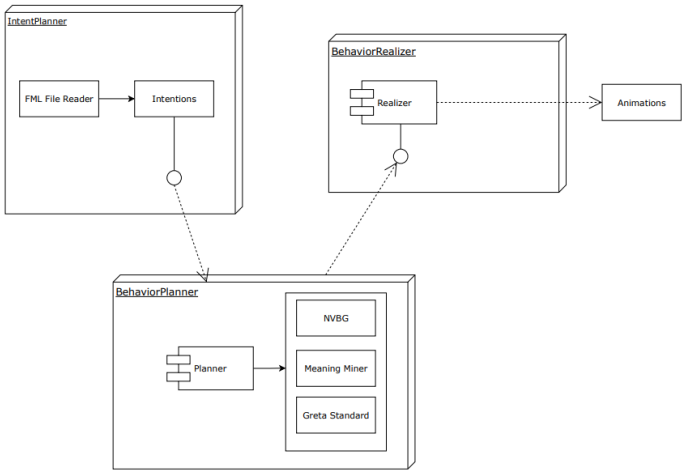

Figure 1: Reviewed BehaviorPlanner Treatement

the agent wants to convey. It allows a virtual agent to display a larger spectrum of behaviors.
In the foreseeable future another model that computes where to place a gesture using deep learning approaches[3, 11]. It will allow Greta to extend its range of gesture, improve gestural performances for life-like virtual characters.

7 ACKNOWLEDGMENT

We thank Philippe Gauthier for his help and support.

REFERENCES

[1] C. Clavel B. Ravenet, C. Pelachaud. 2018. Automatic Nonverbal Behavior Generation from Image Schemas. 9 (July 2018).
[2] Eugene Charniak. 2002. A Maximum-Entropy-Inspired Parser. *Proc NAACL* 1 (05 2002).
[3] Mireille Fares. 2020. Towards Multimodal Human-Like Characteristics and Expressive Visual Prosody in Virtual Agents. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 743–747.
[4] Dirk Heylen, Stefan Kopp, Stacy Marsella, Catherine Pelachaud, and Hannes Vilhjálmsson. 2008. The Next Step towards a Function Markup Language. *International Workshop on Intelligent Virtual Agents*, 270–280. https://doi.org/10.1007/978-3-540-85483-8_28
[5] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn Thórisson, and Hannes Vilhjálmsson. 2006. Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. *Intelligent Virtual Agents, Springer LNCS* 4133, 205–217. https://doi.org/10.1007/11821830_17
[6] Silva Ladewig and Jana Bressem. 2013. *Linguistic perspective on the notation of gesture phases.* 1060–1079.
[7] Jina Lee and Stacy Marsella. 2006. Nonverbal Behavior Generator for Embodied Conversational Agents. 243–255. https://doi.org/10.1007/11821830_20
[8] Isabella Poggi, Catherine Pelachaud, F. Rosis, Valeria Carofiglio, and Berardina Carolis. 2005. *Greta. A Believable Embodied Conversational Agent.* 3–25. https://doi.org/10.1007/1-4020-3051-7_1
[9] Brian Ravenet, Catherine Pelachaud, Chloé Clavel, and Stacy Marsella. 2018. Automating the Production of Communicative Gestures in Embodied Characters. *FRONTIERS IN PSYCHOLOGY* (2018).
[10] Jürgen Streeck. 2013. Elements of Meaning in Gesture, Geneviève Calbris John Benjamins Publishing Company (2011), pp. 378 + VIII. Price: EUR 95.00 | USD 143.00, ISBN: 978-90-272-2847-5. *Lingua* 134 (09 2013). https://doi.org/10.1016/j.lingua.2013.06.005
[11] Fajrian Yunus, Chloé Clavel, and Catherine Pelachaud. 2020. Sequence-to-Sequence Predictive Model: From Prosody To Communicative Gestures. (08 2020).