Perception Markup Language: Towards a Standardized Representation of Perceived Nonverbal Behaviors

Stefan Scherer, Stacy Marsella, Giota Stratou, Yuyu Xu, Fabrizio Morbini,
Alesia Egan, Albert (Skip) Rizzo and Louis-Philippe Morency

University of Southern California
Institute for Creative Technologies, Los Angeles, California
scherer@ict.usc.edu

Abstract. Modern virtual agents require knowledge about their environment, the interaction itself, and their interlocutors' behavior in order to be able to show appropriate nonverbal behavior as well as to adapt dialog policies accordingly. Recent achievements in the area of automatic behavior recognition and understanding can provide information about the interactants' multimodal nonverbal behavior and subsequently their affective states. In this paper, we introduce a perception markup language (PML) which is a first step towards a standardized representation of perceived nonverbal behaviors. PML follows several design concepts, namely *compatibility and synergy*, *modeling uncertainty*, *multiple interpretative layers*, and *extensibility*, in order to maximize its usefulness for the research community. We show how we can successfully integrate PML in a fully automated virtual agent system for healthcare applications.

Keywords: perception, standardization, multimodal behavior analysis, virtual human system

1 Introduction

Human face-to-face communication is a complex bi-directional multimodal phenomenon in which interlocutors continuously emit, perceive and interpret the other person's verbal and nonverbal displays and signals [9, 5]. Interpreting a person's behavior to understand his or her intent requires the perception and integration of a multitude of behavioral cues, comprising spoken words, subtle prosodic changes and simultaneous gestures [13].

Many recent achievements in automatic behavior analysis enable automatic detection, recogniton or prediction of nonverbal behavioral cues, such as laughter [15], voice quality [14], backchannels [8], or gestures [17]. For the rapid advancement of virtual agent systems it is crucial to establish an infrastructure that allows researchers to efficiently integrate these sensing technologies and share their new developments with other researchers. In this paper we introduce perception markup language (PML) as a first step towards a standard representation of perceived nonverbal behavior. The standardization of PML was inspired by

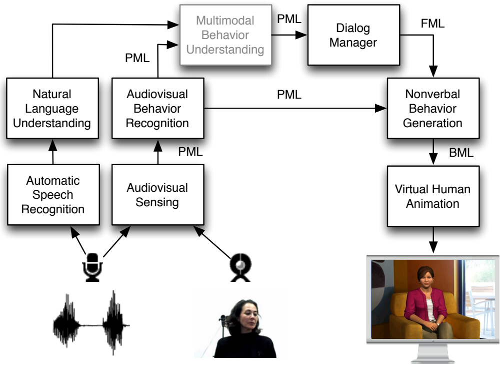

Multimodal Behavior Understanding
PML
Dialog Manager
FML
PML
Natural Language Understanding
Audiovisual Behavior Recognition
PML
Nonverbal Behavior Generation
PML
BML
Automatic Speech Recognition
Audiovisual Sensing
Virtual Human Animation

Fig. 1. Schematic overview showing how perception markup language (PML) can be used in a virtual human architecture.

efforts in the field of nonverbal behavior generation where behavior markup language (BML) and functional markup language (FML) have been introduced in order to enable standardized interfaces for virtual human behavior animation [6, 4].

We show, in Section 3, how PML can interface between sensing and other modules (see Figure 1). PML enables interactive virtual humans to react to the user's nonverbal behaviors. With PML a virtual human system can, for example, provide a wide range of verbal and nonverbal backchannel feedback such as a head nod or para-verbals (e.g., uh-oh) that signal attention, comprehension, (dis-)agreement or emotional reaction to the perceived utterance. This promotes enhanced bi-directional conversations that improve the fidelity of the interaction.

We implemented PML in a real-world healthcare application called "Ellie". Ellie is designed to help detect behaviors related to depression and post-traumatic stress disorder (PTSD) and offers related information if needed. We show in this paper how PML can successfully interface between sensing and higher-level modules such as the dialog manager (DM) and nonverbal behavior generation (NVBG).

2 Perception markup language

In this section we describe our perception markup language (PML), which takes full advantage of the well-established XML standard. We first express the four design concepts behind PML, then give a formal definition through two examples and finally describe PML interaction with DM and NVBG.

2.1 Design concepts

In the following we discuss the main design concepts behind PML.

Compatibility and synergy. Significant effort has been dedicated to building standards for virtual human animation (e.g., FML, BML) [6, 4, 16], speech and language representation (e.g., VoiceXML[1]) and user emotional state (e.g.,

[1] http://www.w3.org/TR/voicexml30/

EmotionML[2]). When designing PML, we carefully analyzed previous standards to learn from their experience and keep PML as compatible as possible. For example, we followed naming conventions of the BML standard whenever possible, as we envision a close interaction between the two standards. Also, instead of reimplementing a textual analysis layer to PML, we plan to work closely with existing speech standards such as VoiceXML. By following these guidelines we not only accelerate the development of a standard, but we also make PML more accessible to the community.

Modeling uncertainty. One of the biggest differentiators between PML and previous standards for virtual human animation (e.g., BML, FML) is the requirement of modeling the inherent uncertainty in sensing and interpreting human nonverbal behaviors. The same visual gesture such as a gaze away can be interpreted as a thinking moment or a disengagement behavior. The language needs to be able to handle multiple hypothesis with their own uncertainty measure. Also, the audio and visual sensing is prone to noise and errors (e.g., due to occlusion or quick movement) which may result in observations with low confidence (i.e., high uncertainty). The correct handling of uncertainty within such modules not only leads to more robust predictions, but might even improve generalization capabilities.

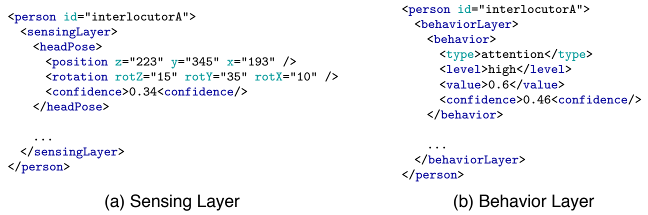
Multiple interpretative layers. When building computational representations of nonverbal communication, people naturally identify multiple layers of interpretation. A concrete example of this is seen in the SAIBA framework [6] which defines separate layers for the behaviors (BML) and their dialog functions (FML). Since the BML is processed by an animation module (e.g., SmartBody) to create the final animation parameters, we can even identify a third layer which includes these animation parameters sent to the realization engine. PML follows the same logic by predefining three layers: sensing, behaviors and functions. Further, these layers allow for the versatile use of PML messages. Some components might solely be interested in higher level observations, while others might analyze rawer data.

Extensibility. Since the field of human behavior recognition and understanding is a constantly growing and developing one, we expect that the XML schema of PML will require multiple updates and revisions even after a deployable version is attained. As technologies develop, the language should develop and adapt to changing requirements. Through collaboration with researchers developing new technologies and researchers using the language, the standard elements of PML will be expanded. The schema can also be automatically converted into code usable for various programming languages following a few processing steps, rendering PML an easily maintainable and extensible markup language.

2.2 Perception markup language specification

Based on these design concepts, we developed the perception markup language (PML) which is a multi-layer representation of perceived nonverbal behaviors and their uncertainty. PML contains two main sections: `<header>` refers to the

[2] http://www.w3.org/TR/emotionml/



```
<person id="interlocutorA">
  <sensingLayer>
    <headPose>
      <position z="223" y="345" x="193" />
      <rotation rotZ="15" rotY="35" rotX="10" />
      <confidence>0.34<confidence/>
    </headPose>

    ...
  </sensingLayer>
</person>
```

(a) Sensing Layer

```
<person id="interlocutorA">
  <behaviorLayer>
    <behavior>
      <type>attention</type>
      <level>high</level>
      <value>0.6</value>
      <confidence>0.46<confidence/>
    </behavior>

    ...
  </behaviorLayer>
</person>
```

(b) Behavior Layer

Fig. 2. PML sample sensing layer (left) and behavior layer (right).

meta-data section (e.g. time stamps and information source) and `<person>` encloses the perceived nonverbal behaviors of a specific person. PML predefines three different layers: `<sensingLayer>`, `<behaviorLayer>` and `<functionLayer>`.

The `sensingLayer` layer provides information from the multiple sensing technologies about the audiovisual state of the users such as their gaze direction, their intonation or their body posture. The `behaviorLayer` layer represents the nonverbal behaviors recognized by integrating temporal information from one or more sensing cues. For example, this layer integrates head and eye gaze to estimate attention behavior or, head and arm motion to estimate fidgeting and rocking behaviors. The `functionLayer` provides information about the user's intent, functional role or affective state of the associated behavior. These higher level concepts are usually estimated by integrating verbal and nonverbal behaviors with contextual information. This paper focuses on the first two layers, keeping the `functionLayer` as future work where we plan to interface this layer with other high-level markup languages such EmotionML and FML. The remainder of this section explains the different parts of a typical PML message.

Message header `<header>`. The header of a PML message includes metadata such as the time stamp that is used for synchronization and a list of datasources `source` identified by a `name` and `id`. The datasource `id` is reflected in the observations within the message `<person>`.

Message body `<person>`. The message body refers to a single person specified with a unique identifier `id`. As discussed, the information associated with each person is separated into multiple layers. The `<sensingLayer>` layer provides information about the current instantaneous audiovisual states. It includes, but is not limited to, fields such as: `gaze, headPose` and `posture`. Each item of the `<sensingLayer>` provides varying information relevant to the field, for example `gaze` provides information on the vertical and horizontal rotation of the eyes and `headPose` provides the coordinates and rotation degrees of the head in the current moment. The `confidence` represents one offered approach to model the uncertainty in sensing technologies. Other fields such as `covariance` can be used to specify the full covariance matrix. An example of the `<sensingLayer>` is seen in Figure 2 (a). The `<behaviorLayer>` layer includes information gathered over longer time periods or inferred complex behaviors. Information such as the attention behavior of the perceived person is transmitted within a `behavior` item. Again the modularity of the approach is seen in the example below, where `behavior` items are structured similarly in order to have an easily extensible approach that can grow with the development of the technology and the demands of connected components. Each `behavior` is identified with a `type`, and the values `level` (categorical; low, mid, high) and `value` (continuous; $\in [0,1]$) indicate the behavior strength. Again, `confidence` indicates the certainty associated with the items as seen in Figure 2 (b).

2.3 PML interaction with other modules

This section describes an example of how PML can interact with dialog management and the nonverbal behavior generation.

Processing PML in DM. The dialogue manager (DM) is tasked to keep track of the state of the conversation, and then decides when and what action to execute next. The DM can select actions that drive the virtual character to say a particular line or execute some specified nonverbal behavior. In this example we use an information state based DM (see [19]) designed to support flexible mixed initiative dialogues and to simplify the authoring of virtual characters. Every time an event is received, to find which action to execute in response, the DM simulates possible future conversations and then selects the action that achieves the highest expected reward. To support relatively high frequency PML events in our virtual agent system, a forward search is not initiated for each PML event that is received, but instead, the message *silently* updates the information state. This in turn affects the action selection procedures of the DM. So, if the audiovisual sensing module identifies a lack of user attention a dialog policy will be triggered to inquire about this possible lack of engagement. Two similar examples are shown in the supplemental video.

Processing PML in NVBG. Whereas, the nonverbal behavior generator (NVBG) [7] automates the generation of physical behaviors for our virtual humans, including nonverbal behaviors accompanying the virtual humans dialog, responses to perceptual events as well as listening behaviors. The handling of PML messages represents a different use case than generating nonverbal behavior for the virtual human's own utterances. In the case of the PML messages, NVBG is deciding on how to respond to perceptual signals about the human's behavior. A human's responses to others' nonverbal behavior, such as mirroring behavior and generic feedback, can in large be measured automatically as opposed to having an explicit communicative intention like an utterance.

Specifically, NVBG's response is determined by a perceptual analysis stage that leads into the behavior analysis and BML generation stages discussed previously. How the virtual human reacts towards actual PML messages in an example interaction is shown in Section 3.2.

3 Use case: Virtual human for healthcare application

The use case scenario in this paper is aimed at subjects and patients interacting with *Ellie*, a virtual human healthcare provider. The interactive sensing system detects depression and PTSD relevant indicators and can offer a faster screening process to a population that often experiences significant wait-times before seeing a real clinician or therapist. The behavioral cues, or indicators include, but are not limited to, behaviors such as lack of expressivity in speech [10, 3], constant and ongoing gaze aversion and lack of mutual gaze [21, 10], as well as increased amounts of anxiety expressed by a rocking motion or fidgeting [2, 11].

3.1 Implementation details

Figure 1 shows the interactional loop of our virtual agent. The *automatic speech recognition* is performed CMU's pocket Sphinx [1] with a push-to-talk approach. The acoustic and language models were trained using transcribed interactions from our pre-study. The recognized utterance is sent to the *natural language understanding* module which recognizes speech acts such as question, statement and backchannel. For the *audiovisual sensing* component we have developed a flexible framework for sensing, based on the social signal interpretation framework (SSI) by [20]. We integrated the following sensing technologies: Cogito Health's *Cogito Social Signal Platform* (CSSP) to extract the speaking fraction as well as other audio features for the speaker, OMRON's *OKAO Vision* for the eye gaze signal, and *CLM FaceTracker* by [12] for facial tracking and head position and orientation, and Microsoft Kinect skeleton tracker.

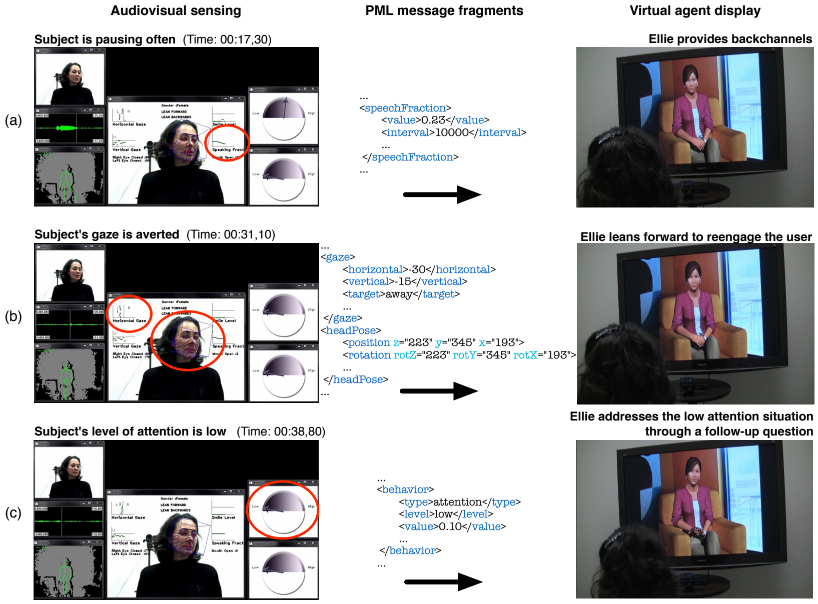
For the *audiovisual behavior recognition* module we implemented memory-driven rule-based system which integrates audio-visual information over time to get higher-level signals such as attention (measured by the gaze signal and face orientation) and activity (measured by body pose information).

The *dialogue manager* has the task to keep track of the state of the conversation and decides when and what action to execute. We use an information state based dialogue manager (see [19]) designed to support flexible mixed initiative dialogues and simplify the authoring of virtual characters. In our scenario, the verbal and nonverbal messages are integrated directly by the dialogue manager instead of having a separate *multimodal behavior understanding* module (as originally shown in the Figure 1). We used the same technique described in Section 2.3 to automatically generate virtual human nonverbal behavior based on the generated utterances (sent by the dialogue manager through the FML messages) and the perception messages (PML). We then use the SmartBody animation module [18] to analyze the BML messages and produce the animation parameters. The final virtual human animations are created using the Unity game engine.

3.2 Example and PML analysis

Figure 3 exemplifies a detailed analysis of a typical interaction with our virtual agent and highlights several key moments when PML messages are used. In these key moments, *Ellie* reacts towards the subject's nonverbal behavior in a way that would not have been possible without the information provided by PML. She for example, exhibits a *head nod* when the subject is pausing a lot in the conversation to encourage the subject to continue speaking (see Figure 3 (a)). In Figure 3 (b), the subject exhibits low attention by looking away. A PML message with this information is sent to NVBG and the virtual agent is signaled to *lean forward*, as an effort to engage the subject. Figure 3 (c) shows an instance where PML signals the DM that the subject's *attention level* is low. This message triggers a branching in the dialog policy[3].

[3] For further information and examples refer to: http://projects.ict.usc.edu/pml


Audiovisual sensing
PML message fragments
Virtual agent display
Subject is pausing often (Time: 00:17,30)
Ellie provides backchannels
(a)
Subject's gaze is averted (Time: 00:31,10)
Ellie leans forward to reengage the user
(b)
Subject's level of attention is low (Time: 00:38,80)
Ellie addresses the low attention situation through a follow-up question
(c)

```
...
<speechFraction>
      <value>0.23</value>
      <interval>10000</interval>
      ...
 </speechFraction>
...
```

```
...
<gaze>
      <horizontal>-30</horizontal>
      <vertical>-15</vertical>
      <target>away</target>
      ...
 </gaze>
<headPose>
      <position z="223" y="345" x="193">
      <rotation rotZ="223" rotY="345" rotX="193">
      ...
 </headPose>
...
```

```
...
<behavior>
      <type>attention</type>
      <level>low</level>
      <value>0.10</value>
      ...
 </behavior>
...
```

Fig. 3. Display of multimodal nonverbal behavior analysis (i.e. Multisense; left column) shown with corresponding PML message fragments (middle column) and virtual agent reactions (right column). Times indicate position in supplementary video.

4 Conclusions

We introduced the perception markup language (PML), a first step towards standardizing perceived nonverbal behaviors. Further, we discussed how the PML messages are used to either change dialog policies or the virtual agent's nonverbal behavior. We provided a detailed walkthrough of a current version of our system with the help of an example interaction, which is provided in full as a supplementary video to the submission of this paper.

In the long run, PML will enable collaborations between currently often isolated workgroups as well as increase the reusability of previous findings and implementations.

Acknowledgements

This work was funded in part by the DARPA Detection and Computational Analysis of Psychological Signals (DCAPS) program and in part by the National Science Foundation grant #IIS-1018055. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

1. *Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices*, volume 1, 2006.

2. L. A. Fairbanks and et al. Nonverbal interaction of patients and therapists during psychiatric interviews. *J. Abnorm. Psychol.*, 91(2):109–119, 1982.
3. J. A. Hall and et al. Nonverbal behavior in clinician-patient interaction. *Appl. Prev. Psychol.*, 4(1):21–37, 1995.
4. D. Heylen and et al. The next step towards a function markup language. In *Proc. of IVA'08*, pages 270–280. Springer, 2008.
5. A. Kendon, editor. *Nonverbal Communication, Interaction, and Gesture.* Number 41 in Selections from Semiotica Series. Walter de Gruyter, 1981.
6. S. Kopp and et al. Towards a common framework for multimodal generation: The behavior markup language. In *Proc. of IVA'06*, pages 21–23, 2006.
7. J. Lee and S. Marsella. Nonverbal behavior generator for embodied conversational agents. In *Proc. of IVA'06*, pages 243–255. Springer, 2006.
8. D. Ozkan and L.-P. Morency. Modeling wisdom of crowds using latent mixture of discriminative experts. In *Proc. of ACL HLT'11*, pages 335–340. Association for Computational Linguistics, 2011.
9. A. Pentland. *Honest Signals - How they shape our world.* MIT Press, 2008.
10. J. E. Perez and R. E. Riggio. *Nonverbal social skills and psychopathology*, pages 17–44. Nonverbal behavior in clinical settings. Oxford University Press, 2003.
11. D. M. Pestonjee and S. C. Pandey. A preliminary study of psychological aftereffects of post-traumatic stress disorder (ptsd) caused by earthquake : the ahmedabad experience. Technical Report WP2001-04-01, Indian Institute of Management, 2001.
12. J.M. Saragih and et al. Face alignment through subspace constrained mean-shifts. In *Proc. of ICCV'09*, pages 1034–1041. IEEE, 2009.
13. S. Scherer and et al. A generic framework for the inference of user states in human computer interaction: How patterns of low level communicational cues support complex affective states. *JMUI, special issue on: Conceptual frameworks for Multimodal Social Signal Processing*, pages 1–25, 2012.
14. S. Scherer and et al. Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. *CSL*, 2012.
15. S. Scherer and et al. Spotting laughter in naturalistic multiparty conversations: a comparison of automatic online and offline approaches using audiovisual data. *ACM TiiS: Special Issue on Affective Interaction in Natural Environments*, 2(1):4:1–4:31, 2012.
16. M. Schröder. *The SEMAINE API: A component integration framework for a naturally interacting and emotionally competent Embodied Conversational Agent.* PhD thesis, Saarland University, 2011.
17. E. A. Suma and et al. Faast : The flexible action and articulated skeleton toolkit. *Virtual Reality*, pages 247–248, 2011.
18. M. Thiebaux and et al. Smartbody: behavior realization for embodied conversational agents. In *Proc. of AAMAS'08*, AAMAS '08, pages 151–158, 2008.
19. D. R. Traum and S. Larsson. The information state approach to dialogue management. In J. Kuppevelt, R. W. Smith, and N. Ide, editors, *Current and New Directions in Discourse and Dialogue*, volume 22, pages 325–353. Springer, 2003.
20. J. Wagner, F. Lingenfelser, N. Bee, and E. André. Social signal interpretation (ssi). *KI - Kuenstliche Intelligenz*, 25:251–256, 2011. 10.1007/s13218-011-0115-x.
21. P. Waxer. Nonverbal cues for depression. *Journal of Abnormal Psychology*, 83(3):319–322, 1974.