Thoughts on FML: Behavior Generation in the Virtual Human Communication Architecture

Jina Lee
University of Southern California
Information Sciences Institute
4676 Admiralty Way, # 1001
Marina del Rey, CA 90292
jinal@isi.edu

David DeVault
USC Institute for Creative Technologies
13274 Fiji Way
Marina del Rey, CA 90292
devault@ict.usc.edu

Stacy Marsella
University of Southern California
Information Sciences Institute
4676 Admiralty Way, # 1001
Marina del Rey, CA 90292
marsella@isi.edu

David Traum
USC Institute for Creative Technologies
13274 Fiji Way
Marina del Rey, CA 90292
traum@ict.usc.edu

ABSTRACT

We discuss our current architecture for the generation of natural language and non-verbal behavior in ICT virtual humans. We draw on our experience developing this architecture to present our current perspective on several issues related to the standardization of FML and to the SAIBA framework more generally. In particular, we discuss our current use, and non-use, of FML-inspired representations in generating natural language, eye gaze, and emotional displays. We also comment on some of the shortcomings of our design as currently implemented.

1. OVERVIEW

In this paper, we discuss our experience developing multimodal generation capabilities within the ICT virtual human architecture. This paper is intended to contribute to an ongoing effort to standardize Functional Markup Language (FML) as a representation scheme for describing communicative and expressive intents across diverse conversational agents. Our discussion focuses on how our current approach to generating natural language, eye gaze, and emotional displays relates to FML and to the SAIBA framework within which FML has been characterized [8].

The SAIBA framework makes a distinction between processes of intention planning, behavior planning, and behavior realization. It then situates these processes within a generation pipeline, and proposes two communication languages to mediate between these processes: FML to specify the result of intention planning to behavior planning, and BML to specify the result of behavior planning to behavior realization.

While there has been a lot of work on BML, there has been comparatively less work on FML and the various real-world architectural issues associated with implementing the SAIBA framework. We begin with a high-level discussion of some of these architectural issues.

One high-level consideration is that the distinction between intention planning, behavior planning, and behavior realization is only one of many organizing distinctions that could be made in a communication/action planning framework. Some others include the following.

One can distinguish actions according to the different kinds of intentions that can be behind them. Allwood [1] distinguishes three types of communication: Indicate, Display, and Signal. A sender *indicates* information if that information is conveyed without conscious intention. *Displays* are consciously shown, and *signals* are conscious showings of the showing (i.e. intending the receiver to recognize the conscious showing). An embodied agent may perform an action intentionally without intending to communicate anything; if another agent or person is present, important information may nevertheless be conveyed by indication. Should the planning of actions that are not intended to be communicative be part of the FML/BML pathway, or should these actions reach the behavior realizer through some other channel? Moreover, some behaviors that embodied agents need to realize (e.g., breathing) are not "intentional" in the relevant sense, and thus the notion of intention planning is inappropriate. If information about agent state is relevant to realizing such behaviors, is this information also channeled to the realizer outside the FML/BML pathway?

Another organizing distinction could be the type of behavior. Traditionally, verbal behavior and non-verbal behavior have been generated at different times and using different means. Verbal communication has discrete units, a fairly arbitrary relationship of form to meaning, and deep lexical, syntactic and semantic structures, while non-verbal communication often is more continuous, has a closer relationship of form to meaning, and shallow syntactic structure. Traditional text generation often has more stages in processing, and uses more contextual information. Most SAIBA work has focused on non-verbal behavior. Should the same pathways be used for text generation and non-verbal behavior, or should these paths be split (e.g., with text generated first)? And of course, this issue extends to other kinds of behaviors that are not realizing a communicative function.

Another architectural issue arises in real-time interactive considerations. Even though the proponents of the SAIBA framework are keenly aware of the importance of real-time interaction, the SAIBA framework remains suggestive of a traditional pipeline architecture of planning followed immediately by plan execution. This is fine for a virtual agent that resides in a static environment. However, in a more dynamic environment, an agent must respond to unexpected events in the environment. For example, many communication decisions must rely not just on individual intention planning, but also on monitoring the effects of previously planned action, and especially on monitoring new actions by people and other agents. Intention planning thus must have access to this information and must also be able to adjust or cancel communication that has been planned but not yet performed. This suggests not only additional requirements on what is provided by the intention planner to the behavior planner but also on what is provided by the behavior planner and realizer to the intention planning.

Finally, there is a more general architectural question of how to modularize a real-world generation system in a way that provides each module with all the sources of information it needs. For example, as we discuss in further detail below, our current gaze generation system relies on fine-grained, dynamic information about upstream cognitive processing. Similarly, natural language generation can sometimes require detailed information about the agent's cognitive state and other contextual factors. Such rich information needs can create pressures that work against maintaining a clean theoretical modularity such as that suggested in the SAIBA framework.

In the remainder of the paper, we discuss our virtual human architecture and then our perspective on how our current design might inform the standardization of FML.

2. ICT VIRTUAL HUMAN COMMUNICATION ARCHITECTURE

The virtual human project at ICT [14, 20, 17] has produced several virtual humans and a developing architecture, which is depicted in Figure 1. In this section, we describe the control flow and representations involved in generating multimodal output within this architecture.

For intentional communication signals, the generation process starts with configurations of the agent's information state that match a *proposal rule*. Examples include obligations to answer a question, ground or repair previously communicated information, or make a suggestion. These proposals to communicate compete with many other goals of the agent – both to say other things as well as to perform other actions such as monitoring the communication of others or acting in the world. Once a proposal is selected, the generation process begins.

2.1 Natural language generation

In our current system, natural language generation (NLG) occurs before non-verbal behavior generation (NVBG). In general, the dialogue manager initiates NLG by sending a generation request to an external generator. However, currently the dialogue manager sometimes bypasses the external generator if it already knows a good text string for its desired output, according to hand-implemented SOAR rules, or rules generated from an ontology. We have four different external generators that may be used, including two statistical generators, a hand-crafted grammar-based generator, and a hybrid generator. [19] has more details on a previous version of the generation process.

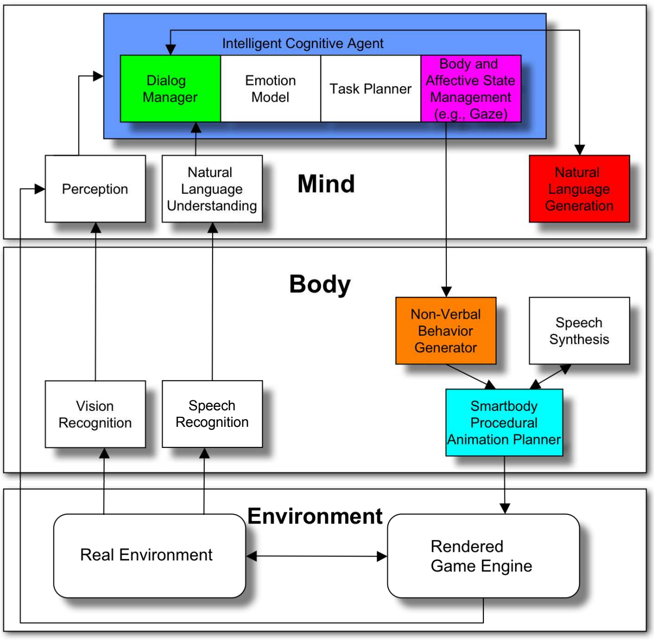
Intelligent Cognitive Agent
Dialog Manager
Emotion Model
Task Planner
Body and Affective State Management (e.g., Gaze)
Perception
Natural Language Understanding
Mind
Natural Language Generation
Body
Non-Verbal Behavior Generator
Speech Synthesis
Vision Recognition
Speech Recognition
Smartbody Procedural Animation Planner
Environment
Real Environment
Rendered Game Engine

Figure 1: The virtual human system architecture.

The dialogue manager sends requests to the generator in the form of one or more speech acts and dialogue acts to realize. The messages to the generator are of the form given in Figure 2. The `vrGenerate` message can be received by any external generator. In this case the dialogue manager is asking for a `greeting` speech act from the virtual human `elder-al-hassan` to a human addressee, who plays the role of a U.S. Army captain (`captain`). This act is also the response to a previous utterance. One or more generators can reply to this request with `vrGeneration` messages such as those in Figure 3. There can be one or more `vrGeneration interp` messages, each one with a candidate text for this output and with an interpretation identifier (`1`) and a quality value (`-3.742008`). The `vrGeneration done` message tells the dialogue manager that the generator(s) are finished sending interpretations.

Figures 4 and 5 show a similar request and response. This time another virtual human, `doctor-perez`, is trying to negotiate, and wants to address a problem in a plan involving moving to downtown by telling the elder that his agreement is important for the success of the plan. When the dialogue manager has received the generation results, it can decide which one to use (if there is more than one result), based on both the quality of the generation and other factors (e.g., whether it has said this same string before). The dialogue manager might also decide to cancel the speech if it is no longer relevant (or if, e.g., another character starts speaking and this character does not want to interrupt).

Thus, in our current architecture, NLG is not part of a pure pipeline since the upstream dialogue manager chooses between alternative NLG outputs and sometimes cancels output altogether. After the dialogue manager decides to go forward, a call is sent to carry out this utterance. This call includes information on the speech acts and dialogue acts as well as the text, and results in an XML message

```
vrGenerate elder-al-hassan elder-al-hassan203
  addressee captain
  speech-act<A135>.type csa
  speech-act<A135>.action greeting
  speech-act<A135>.actor elder-al-hassan
  speech-act<A135>.response-to gsym1
  speech-act<A135>.addressee captain
```

Figure 2: Generator request

```
vrGeneration interp elder-al-hassan
             elder-al-hassan203 1 -3.742008
             hello captain
vrGeneration done elder-al-hassan
             elder-al-hassan203
```

Figure 3: Generator response

being sent to the NVBG module.

2.2 Nonverbal and other physical behaviors

In addition to dialogue management, a virtual human's cognitive processes include task planning, a gaze model, and an appraisal-based model of emotion. These processes provide a range of information to our NVBG module [10] through FML-*inspired* constructs. This information includes a specification of the communicative intent (including the speech acts and dialogue acts), the surface text of the utterance, the agent's gaze state, and a range of factors associated with the emotion model.

In this section, we present our current use of FML-inspired constructs to pass gaze and emotion information to the behavior planner. We will not discuss further the simple FML elements we currently use to capture the communicative intent and the surface text. It is important to note, however, that this is a hybrid approach that assumes NLG is upstream of the behavior planner but also assumes the intentional/semantic content can help refine non-verbal behavior choices. In terms of the SAIBA framework, one way to view this approach is that in some implementations both FML elements and BML elements are passed to the behavior planner. More generally, this raises fundamental issues for FML and SAIBA as to what assumptions are being made in the framework about how verbal and nonverbal behaviors are generated (or co-generated). We discuss this in greater detail in Section 3.3. Presently, we are actively considering alternative generation schemes and therefore expect our perspective on the appropriate FML elements to evolve as our design process continues. Our focus in this section is on aspects of our current use of FML that are somewhat more stable and, we believe, more transferable to other systems.

2.2.1 Gaze

The reader may think that gaze is not a function but a behavior, and thus should not be an element in FML at all, but rather solely in BML. In the abstract, we would tend to agree. However, given the real-time changes in human gaze directions and targets during communication, and the myriad functions that gaze plays in human cognitive and social behavior, it is important to consider its role in detail.

```
vrGenerate doctor-perez doctor-perez386
  addressee elder-al-hassan
  speech-act<A348>.motivation<V22>.reason
                                    downtown
  speech-act<A348>.motivation<V22>.goal
                                    address-problem
  speech-act<A348>.content<V21>.
           modality<V23>.conditional should
  speech-act<A348>.content<V21>.type action
  speech-act<A348>.content<V21>.theme downtown
  speech-act<A348>.content<V21>.event agree
  speech-act<A348>.content<V21>.agent
                                  elder-al-hassan
  speech-act<A348>.content<V21>.time present
  speech-act<A348>.addressee elder-al-hassan
  speech-act<A348>.action assert
  speech-act<A348>.actor doctor-perez
```

Figure 4: Generator request

```
vrGeneration interp doctor-perez
    doctor-perez386 1 -2.9832053
    you should agree to this before we can think
    about moving elder
vrGeneration done doctor-perez doctor-perez386
```

Figure 5: Generator response

In our current virtual human system, the gaze model [11] resides in the cognitive module and generates various gaze commands. The key principle behind the model is that gaze should reflect the agent's underlying cognitive state; this has historically led us to locate it within the cognitive module, not the behavior planner. Since gaze movement is a fast and immediate process, the gaze model is closely intertwined with the agent's task planner, dialog manager, and emotion model. Each of these components, which constitute the cognitive module, generates a set of cognitive operators that represent the agent's internal processing. The role of the gaze model is then to associate these operators with corresponding gaze behaviors.

The generated cognitive operators can be understood in terms of several broad categories of cognitive processes in conversation. For example, as illustrated in Table 1, there are cognitive operators related to conversation regulation, update of internal cognitive state, and monitoring of events or goal status. While most operators related to conversation regulation generate gaze commands accompanying verbal utterances, others do not. For instance, monitoring for expected/unexpected changes, attending to a physical stimulant in the environment, or checking a condition for a pursued goal are internal intentions that are reflected intentionally or unintentionally through various nonverbal behaviors. Additionally, there are cognitive operators related to the agent's coping strategies (discussed further below).

The gaze model associates these cognitive operators with gaze behaviors by providing a specification of *both* the physical manner of gaze (e.g. target, type, speed, priority) and its functional role. The functional role, or the *reason* of the gaze command, is a description of the cognitive operator that triggers the gaze command. This may be a sub-phase of a higher-level cognitive operator. For example, during

Category	Cognitive Operator	Gaze Reason
Conversation Regulation	output-speech	- planning_speech_(look_at_hearer, hold_turn, rejection, rejection_goal_satisfied, acceptance_reluctant, remembering) - speaking - speech_done - speech_done_hold_turn
	listen-to-speaker	- listen_to_speaker
	interpret-speech	- interpret_speech
	expect-speech	- expect_speech
	wait-for-grounding	- expect_(acknowledgment, expect_repair)
Update Internal Cognitive State	update-desire update-relevance update-intention	- planning
	update-belief	- monitor_goal
Monitor for Events / Goal Status	attend-to-sound	- attend_to_sound_object
	check-goal-status	- monitor_goal
	monitor-goal-status	- monitor_goal_refresh
	monitor-for-expected-effect	- monitor_for_expected_effect
	monitor-for-expected-action	- monitor_expected_action
Coping Strategy	Coping-focus	- monitor_expected_action (assert intention to perform the action, (take action against an action) - seek_social_support - monitor_goal - avoidance - convey_displeasure - accept_responsibility - make_amends - resignation - avoidance (by-distancing, by-wishing-away)

Table 1: Partial overview of cognitive operators, gaze reasons, and gaze behaviors

the output-speech phase, there are sub-phases such as planning speech, speaking, complete speaking, holding the turn, etc. Table 1 shows how various gaze reasons correspond to cognitive operators in our system.

In our system, we use an FML <gaze> element with the properties of gaze behaviors specified in the attributes and send it to NVBG. NVBG then transforms it into a BML <gaze> element and sends it to SmartBody [18], the behavior realization module.

As the gaze model was originally developed, the gaze manner specified by the model provided parameters to a procedural animation of gaze by a behavior realizer. However, in our current work, we are providing the *reason* parameter to the behavior planner. This specification will allow for more expressive variations as well as variations that may also be tied to other aspects of the body's state as well as capabilities of the animation system.

2.2.2 *Emotion*

In our system, we model both the generation of emotional states that arise as the virtual human reacts to events as well as how the virtual human copes as it attempts to regulate its emotional state. EMA (**EM**otion and **A**daptation) [7] is the emotion model in our virtual human system. EMA is largely based on Lazarus' work on appraisal theory [9].

Appraisal

EMA assesses emotion-eliciting events into a range of appraisal dimensions (or checks or variables), such as perspective, desirability, likelihood, expectedness, causal attribution, temporal status, controllability, and changeability. The appraisal dimension is then mapped to generate various emotion labels and intensity of those emotions. For example, an undesirable and uncontrollable future state is mapped as fear-eliciting. In general, a set of appraisal patterns can generate one or more emotion labels.

Currently in our system, an FML <affect> element is used to specify both the emotion labels along with the intensity, target, and stance (leaked or intended) of the emotion. Whenever the agent's emotion is re-assessed, this information is sent to the NVBG module, which uses it to modify the gestures created. Note we discuss in this section how we model "leaked" emotions, or more accurately "felt" emotions, as opposed to emotional expression used intentionally as a signal, which we discuss in the Coping Strategy section below.

Once the appraisal dimensions are (re-)evaluated, they are also used to generate Facial Action Unit codes, based on the work of Ekman [5]. As opposed to emotion labels, the action units are specified in BML (instead of FML) within the <face> element and sent to NVBG. Since NVBG receives the action units in BML, it simply passes them to SmartBody. However, conceptually it should be the behavior planner that generates action units along with other gestures after receiving the agent's affective state. In the future, we suggest alternative ways to express the agent's affect depending on the level of detail available. Section 3 describes our proposed FML specifications.

Note that there is a range of research issues concerning the mapping from appraisals and emotions to action units that we are glossing over here. Whereas several psychological theories have postulated a mapping from appraisal variables to action units, they differ on the specifics of the mapping. Further, given any specific appraisal, there may not be a unique mapping to action units even given the same theory. There are individual differences in how to map appraisals or emotions to action units. There are also alternative theories that postulate that there is not a mapping between appraisals to action units but rather mappings from emotions to action units. There are also issues in dynamics. Psychological theories differ in whether they postulate temporal ordering relations between appraisal checks and whether they argue that this ordering is reflected in temporal differences in the ordering of associated action units. There are, finally, even some psychologists that argue against facial expressions revealing ”true” underlying emotional states, instead arguing that facial expressions are social signals.

Coping Strategy

EMA also incorporates a computational model for coping strategy integrated with appraisal dimensions [7]. EMA analyzes the causality of events that produce the given appraisal dimensions and suggests strategies to either preserve desirable states or overturn undesirable states. These strategies may propose to execute certain plans, alter goals and beliefs, or shift blame for an undesirable event to another entity. The coping strategies modeled in EMA are organized by their impact on the agent's focus of attention, beliefs, desires, or intentions. Table 2 gives an overview of the coping strategies.

In the current virtual human system, coping strategies are propagated to the behavior planner in two ways. One is by implicitly influencing the agent's affective state and generating a new emotion label, which is then taken into account during behavior generation. The other is by directly influencing the nonverbal behaviors generated. In particular, the gaze model described above has certain gaze behaviors associated with different coping strategies. For example, *seek instrumental support* shifts gaze towards some other agent whereas *Resignation* causes the agent to avert gaze from its current target. However, as with the case of appraisal, it is more appropriate to describe the coping strategy within FML and let the behavior planner decide how this would influence the behavior generation process.

Coping also provides the agent with the means to convey emotional states intentionally, for example, by showing displeasure or anger. This expression or signaling of emotional state may differ from the true or felt underlying emotional state of the virtual human. It is this distinction which motivated the original FML ideas of distinguishing “leaked” from “intended” emotions; see our proposed FML <affect> element in Section 3.2.

Currently, modeling of coping strategies is not common in virtual human systems. Unlike other cognitive operations described in this paper, coping strategies *may* not have an immediate effect in the behavior generation process. Rather a coping response may influence how the agent selects, plans, and executes its internal goals. This in turn has influences on the choices of behaviors. On the other hand, a coping response can be an immediate reaction with well-defined behavioral correlates, such as avoidance responses impacting gaze or shifting blame impacting an expression of anger.

3. PROPOSED SPECIFICATIONS OF FML

In this section, we propose several elements of FML based on our current provisional use of FML-inspired constructs.

3.1 Gaze

As described in the previous section, the key principle in our model of gaze is that it should reflect the agent's inner processing. In line with this, our current and proposed specification of the <gaze> element in FML includes the *reason* of the gaze command in fine-grained detail along with the target and type of gaze (see Table 3). This allows different behavior planners to represent the same communicative intent with varying expressivity depending on the capability of the virtual human system (e.g. full human embodiment vs. simplified character with only a head figure).

A second alternative is to back away from the commitment that the link from cognitive processes to behavior planner is captured solely in FML and the link from behavior planner to realizer is captured solely in BML. Rather, various modules along a path (or paths) may be allowed to add FML or BML elements. This allows for considerable flexibility in how modules are realized but may also impact the sharing of modules across research efforts.

Finally, we could go even further towards a functional specification. FML may want to avoid even calling this element ‘gaze’. Perhaps ‘attention’? However that also does not quite capture the range of functions that is performed by different gaze types. That range might be best expressed by the general categories in Table 1: Conversation Regulation, Update Internal Cognitive State, Monitor, and Coping Strategy. In this view, the FML element would be one of those categories, with the Reason being a further specialization of that element. We believe this view is most consistent with the goals of specifying FML.

3.2 Emotion

Our proposal for representing emotion in FML is to have alternative ways to express the agent's affect. These alternative ways would be tied to the underlying class of emotion model used by a system. For instance, we suggest an FML structure that allows the system to either represent the emotion labels (categories) or the more detailed appraisal dimensions. Table 4 gives the suggested structure of two FML elements for this purpose. Here are examples of both cases:

1. Representing emotional label:
```
<affect type=”joy” intensity=”1.0” target=”captain-kirk”
/>
```

2. Representing appraisal dimensions:
```
<affect type=”appraisals” target=”captain-kirk”/>
  <appraisal type=”desirability” value=”0.2” />
  <appraisal type=”controllability” value=”0.5” />
  ...
</affect>
```

In the latter case, if the value of the affective type is ‘appraisals’, the type, target, and stance of the emotion should still be specified. But we propose the <affect> element to have an arbitrary number of <appraisal> elements embeded to represent the different appraisal variables and values.

Table 2: Coping strategies modeled in EMA

Coping Strategy	**Description**
	Attention Related
Seek Information	Form a positive intention to monitor pending unexpected, or uncertain state that produced the appraisal values.
Suppress Information	Form a negative intention to monitor the pending, unexpected or uncertain state that produced the appraisal values.
	Belief Related
Shift Responsibility	Shift a causal attribution of blame/credit from/towards self and towards/from other agent.
Wishful Thinking	Increase/lower probability of a pending desirable/undesirable outcome or assume some intervening act/actor will improve desirability.
	Desire Related
Distance/Mental Disengagement	Lower utility attributed to a desired, but threatened state.
Positive Reinterpretation / Silver Lining	Increase utility of positive side-effect of some action with a negative outcome.
	Intention Related
Planning / Action Selection	Form an intention to perform some external action that improves an appraised negative outcome.
Seek Instrumental Support	Form an intention to get some other agent to perform an external action that changes the agent-environment relationship.
Make Amends	Form an intention to redress a wrong.
Procrastination	Defer an intention to some time in the future.
Resignation	Abandon an intention to achieve a desired state.
Avoidance	Take action that attempts to remove agent from a looming threat.

Table 3: Proposed structure of <gaze> element in FML

Element:	**<gaze>**
gaze-type	A symbol describing the type of gaze at the target (e.g. avert, cursory, look, focus, weak-focus).
target	The name of an object that the agent is gazing at or shifting gaze to, or averting in the case of gaze aversion.
priority	A symbol describing the priority of the cognitive operation that triggered this gaze command.
reason	A detailed rationale behind why we are doing the gaze (currently represented as a token).

Table 4: Proposed structure of <affect> element in FML

Element:	**<affect>**
type	Indicates the category of affect (joy, anger, fear, ...) or whether the affect will be represented by appraisal dimension (appraisals).
target	Person who is possibly being targeted by the resulting affective behavior.
stance	Whether the emotion is intentionally given off or involuntarily leaked (intended, leaked).
intensity	The intensity of emotion.

Element:	**<appraisal>**
type	A single appraisal variable (desirability, controllability, ...).
value	The intensity of the appraisal variable.

As discussed above, researchers have developed a number of theories of emotions, each varying in how they model the dynamics of emotional processes. Here we have suggested two ways to represent emotion from two emotion theories, namely the categorial theory of emotion and appraisal theory. The expressivity to represent not only the emotion labels but also the appraisal variables allows the behavior planner to draw on a deeper understanding of the impact an event has for an agent and to generate behaviors accordingly. However, to employ models of other emotion theories, more discussion is needed about how to represent the properties of those models. In particular, we should also consider incorporating dimensional models such Mehrabian and Russell's PAD (Pleasure-Arousal-Dominance) model [12] or more recent work related to such dimensional models (e.g., Core Affect). Finally, we should also explore the emotion annotation schemes being developed by other consortiums such as the HUMAINE work [15].

3.3 Language Generation and FML

As discussed in Section 2.1, we currently use a system-specific representation scheme to formulate NLG requests and responses. We have not attempted to transform this scheme into an FML representation that might be used across different systems. In this section, we discuss some of the challenges we believe would be associated with standardizing a messaging protocol for NLG across systems.

In general, our perspective is that if NLG is to be assimilated into the SAIBA framework, it should be viewed as part of behavior planning rather than intent planning. This is because, first, at a conceptual level, language use *is* planned behavior. Indeed, NLG systems typically frame their language generation problem as one of planning a linguistic output that accomplishes an incoming communicative intention or communicative goal [13]. Second, in many systems, there may be advantages in terms of naturalness and efficiency of communication that come with planning verbal and non-verbal behavior *simultaneously*, as in, for example, [2].

Let us consider, then, what the implications for the standardization of FML would be if NLG were to be generally situated within the behavior planning stage of the SAIBA framework. In the canonical NLG pipeline [13], an NLG algorithm is internally divided into three successive stages: document planning, microplanning, and realization. Document planning is the process of deciding what information should be communicated, while microplanning and realization plan an output text that achieves this communicative goal. An intuitive approach would therefore locate document planning within the intent planning stage of SAIBA, and locate microplanning and realization within the behavior planning stage.

To understand the implications for FML, we need to look at the typical inputs needed by microplanners and realizers. While the division of labor between microplanning and realization, and the interface between them, varies considerably between systems [13], we may generally observe that both processes depend on relatively rich input specifications to achieve high quality output. For example, one subtask that microplanners typically solve is the generation of referring expressions (GRE) for particular objects or individuals that are implicated in the communicative goal. In general, GRE requires as input a ranking of the relative salience of various objects and properties in the non-linguistic context, as well as the dialogue/discourse history, so that an appropriate level of detail can be selected for the referent of the expression (e.g., the choice of a pronoun versus a complex definite noun phrase); see, e.g., [3, 16].

More generally, the fact that microplanning and realization involve fine-grained lexical choices can add additional input requirements. For example, the SPUD microplanner [16] requires as input the communicative goal (expressed as a set of logical formulas), a grammar, and a representation of the current context (including elements of dialogue/discourse history as well as non-linguistic context). Because SPUD expects the communicative goal to be expressed using logical formulas, it would not be trivial to translate a virtual human generation request such as those in Figures 2 and 4 into a communicative goal for SPUD. Further, the input context representation needs to extend down to the granularity of lexical semantics in the language to be generated. One way of providing this information to SPUD is to provide a *knowledge interface*, as in [4]. The knowledge interface allows SPUD to interactively query for salience information and to evaluate semantic constraints associated with alternative lexical choices in the current context. This creates another question about how to provide, within the SAIBA framework, an NLG module with all the resources it potentially needs. It would seem that an FML-ized generation request would either need to carry a quite exhaustive description of context, or else the generator would need to be provided with some mechanism by which upstream modules can be interactively queried for additional information as needed.

Another challenge is that different realizers can also expect different input formats. For example, the FUF realizer [6] requires as input a *functional description*, which is a hierarchical set of attribute-value pairs that partially specify the lexico-syntactic structure of the output utterance. The OpenCCG realizer [21] requires as input the logical form of the utterance to be realized, expressed (in XML) as a semantic dependency graph or (equivalently) in a hybrid logic dependency semantics formalism. Typically, for a given realizer, a paired microplanner draws on a lexicon and/or grammar, as well as various domain-specific rules and context information, to automatically translate a communicative goal into the appropriate inputs to the realizer. The challenge for FML is that the particular representation scheme that is chosen for FML should aim to remain compatible with, and easily converted into, the particular input formats and internal pipelines assumed by such different NLG components. We do not immediately see how to achieve this goal, especially given the widely varying approaches to NLG that are currently being explored. However, this is an area where detailed discussion between researchers might yield an operational interim approach.

4. CONCLUSION

In this paper we have presented our implementation of multimodal generation capabilities in the ICT virtual human architecture. We have drawn on our experience with this architecture to present our perspective on the standardization of FML elements for generating eye gaze, emotional displays, and natural language. While our conclusions have generally been tentative, we hope to have achieved our aim of furthering the ongoing discussion of FML and the SAIBA framework as a useful approach to multimodal generation across diverse conversational agents.

5. ACKNOWLEDGMENTS

This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM), and the content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

6. REFERENCES

[1] J. Allwood. Bodily communication - dimensions of expression and content. In B. Granström, D. House, and I. Karlsson, editors, *Multimodality in Language and Speech Systems*, pages 7–26. Kluwer Academic Publishers.

[2] J. Cassell, M. Stone, and H. Yan. Coordination and context-dependence in the generation of embodied conversation. In *Proceedings of INLG*, 2000.

[3] R. Dale and E. Reiter. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263, 1995.

[4] D. DeVault, C. Rich, and C. L. Sidner. Natural language generation and discourse context: Computing distractor sets from the focus stack. In *Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*, pages 887–892, 2004.

[5] P. Ekman and W. Friesen. *The Facial Action Coding System (FACS): A technique for the measurement of facial action.* Consulting Psychologists Press, Palo Alto, CA, USA, 1978.

[6] M. Elhadad. FUF: the universal unifier user manual version 5.0. Technical Report CUCS-038-91, 1991.

[7] J. Gratch and S. Marsella. A domain-independent framework for modeling emotion. *Cognitive Systems Research*, 5(4):269–306, 2004.

[8] S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, and H. H. Vilhjálmsson. Towards a common framework for multimodal generation: The behavior markup language. In *IVA*, pages 205–217, 2006.

[9] R. Lazarus. *Emotion and Adaptation.* Oxford University Press, New York, NY, USA, 2000.

[10] J. Lee and S. Marsella. Nonverbal behavior generator for embodied conversational agents. In *Proceedings of the 5th International Conference on Intelligent Virtual Agents*, 2006.

[11] J. Lee, S. Marsella, J. Gratch, and B. Lance. The rickel gaze model: A window on the mind of a virtual human. In *Proceedings of the 6th International Conference on Intelligent Virtual Agents*, 2007.

[12] A. Mehrabian and J. A. Russell. *An approach to environmental psychology.* MIT Press, Cambridge, MA, USA; London, UK, 1974.

[13] E. Reiter and R. Dale. *Building Natural Language Generation Systems.* Cambridge University Press, New York, NY, USA, 2000.

[14] J. Rickel, S. Marsella, J. Gratch, R. Hill, D. Traum, and W. Swartout. Toward a new generation of virtual humans for interactive experiences. *IEEE Intelligent Systems*, 17:32–38, 2002.

[15] M. Schröder, L. Devillers, K. Karpouzis, J.-C. Martin, C. Pelachaud, C. Peter, H. Pirker, B. Schuller, J. Tao, and I. Wilson. What should a generic emotion markup language be able to represent? In *Proc. 2nd International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 440–451, 2007.

[16] M. Stone, C. Doran, B. Webber, T. Bleam, and M. Palmer. Microplanning with communicative intentions: the spud system. *Computational Intelligence*, 19(4):314–381, 2003.

[17] W. R. Swartout, J. Gratch, R. W. Hill, E. H. Hovy, S. Marsella, J. Rickel, and D. R. Traum. Toward virtual humans. *AI Magazine*, 27(2):96–108, 2006.

[18] M. Thiebaux, A. Marshall, S. Marsella, and M. Kallmann. Smartbody: Behavior realization for embodied conversational agents. In *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems*, To appear.

[19] D. Traum, M. Fleischman, and E. Hovy. Nl generation for virtual humans in a complex social environment. In *Working Notes AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue*, March 2003.

[20] D. Traum, W. Swartout, S. Marsella, and J. Gratch. Virtual humans for non-team interaction training. In *In proceedings of the AAMAS Workshop on Creating Bonds with Embodied Conversational Agents*, July 2005.

[21] M. White, R. Rajkumar, and S. Martin. Towards broad coverage surface realization with ccg. In *Proc. of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+MT)*, 2007.