

Functional Description of Multimodal Acts: A Proposal

Kristinn R. Thórisson
CADIA / School of Computer Science
Reykjavik University
Kringlunni 1, 103 Reykjavik, Iceland
thorisson@ru.is

Hannes H. Vilhjalmsón
CADIA / School of Computer Science
Reykjavik University
Kringlunni 1, 103 Reykjavik, Iceland
hannes@ru.is

ABSTRACT

Architectures for controlling communicative humanoids have been many and varied. Planning systems for multimodal behavior still require significant efforts to design and implement; this could be alleviated to some extent through the use of a common platform. In this paper we outline an approach to multimodal action generation following the SAIBA framework. The proposal focuses on planning at a medium-level of action abstraction – what we refer to as a *functional* level – building on our prior efforts in creating systems capable of human-like multimodal behavior. Starting from a high-level initial goal or releasing mechanism we assume that surface behavior can be generated in continuous incremental steps at multiple levels of abstraction, as outlined by SAIBA, progressing over time in a depth-first manner, towards an actual executed behavior. The paper proposes a starting point for a language intended to represent/describe functional aspects of communicative action. We argue that among key aspects that must be addressed for this to be successful are temporal constraints, prioritization and classification of function.

Categories and Subject Descriptors

D.3.0 [Programming Languages]: General – Standards. D.3.2 [Programming Languages]: Language Classifications – Very high-level languages.

General Terms

Design, Standardization, Languages, Theory

Keywords

Multimodal Communication, Realtime, Communicative Humanoids, Functional Markup Language, SAIBA, Multimodal Acts, Embodied Agents.

1. INTRODUCTION

The SAIBA framework [1] is motivated by a need to enable collaboration in building communicative humanoids. Second, it is motivated by push towards more sophisticated multimodal communicative planning, and third, by a hope for easier construction of multimodal skills for multimodal characters, whether physical robots or virtual. Towards this end, SAIBA proposes a modular approach to the “planning pipeline”¹ for

realtime multimodal behavior. There are at least two important modular splits in this respect. The first is between a representation *language that describes* an action/set of actions and *the engine/mechanism that realizes* these, according to a specification written in this language. Another split – or set of splits – proposed by SAIBA is between lowest-level behaviors (“animation level”), a medium-level representation typically called “behavior” level, and a higher level called the “functional” level. These levels correspond roughly to what have sometimes been called the *primitive/servo level*, *e-move level* and *task level*, respectively, in the robotics community [2].

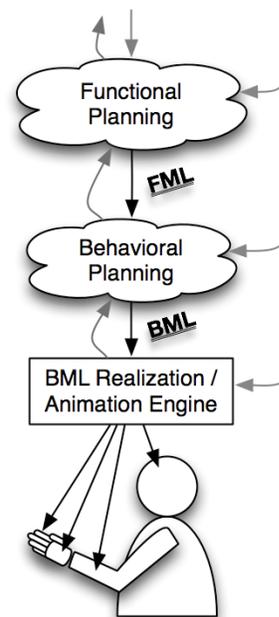


Figure 1: The planning levels envisioned by the SAIBA framework, showing where FML and BML fit in. Upwards gray arrows indicate feedback; gray side arrows imply that other input to the planning mechanisms could come from elsewhere in the system.

At the first split, SAIBA proposes the Behavior Markup Language (BML) as the representation language for describing human movement at the level of behaviors [1][3]. BML can thus serve as

¹ As we assume ample feedback loops in this system via perceptual mechanisms it is, strictly speaking, not a pipeline. This is an important

point that often is overlooked. However, the pipeline model works reasonably as a first approximation.

the input to a basic animation engine. An example of an engine that can realize BML as realtime-executed multimodal actions is SmartBody [4], which we have incorporated into the relatively simple-to-use CADIA BML Realizer².

The idea behind the split between representation language and realization engine is to enable those researchers who desire to focus on a particular level of planning to stick to a certain level of detail. The language describing the desired outcome at a particular detail level can be represented in a common way between researchers, making easier the collaboration on – and competition between – proposed mechanisms. This allows construction of alternative planning mechanisms at particular levels of abstraction, and thus exploration of different ways of producing certain behavior phenomena, without having to solve the mechanism for the whole field, as the representational languages provide an API that allows modular sharing of solutions for different parts of the architecture. This also enables the comparison between realization mechanisms from different research labs. Further benefits to such a modular scheme are discussed in [1][5][6].

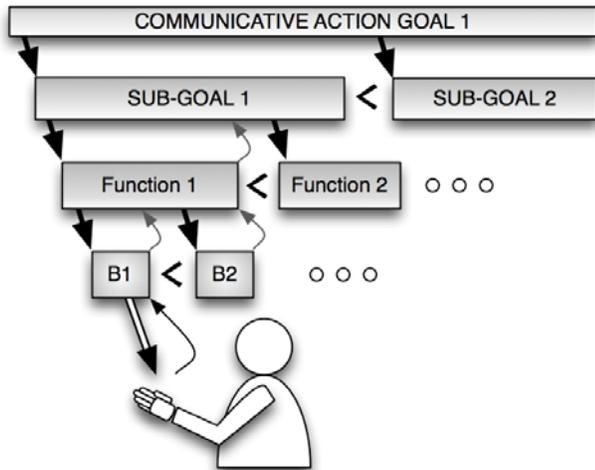


Figure 2: Successive refinement of goals (spanning long durations of time) into functional descriptions (fat arrows) and then into behaviors that can be executed (hollow arrow) as actual movements. Only sequential dependencies between actions are depicted here. Feedback about the actual implementation of each chunk is provided to the next level up (narrow bent arrows).

Looking at things from a descriptive, static perspective, BML describes human multimodal behavior at a particular level of abstraction. For runtime systems BML provides a “human-readable description level”, which is assumed to be the output of a behavior-level engine. Several prototypes of such engines exist in our research community. The *input* to such an engine, however, has not been specified in the context of the SAIBA framework/consortium. SAIBA proposes that the input to this (abstract) engine should be in the form of a language that represents functional aspects of the movements. The idea here is that this could be captured in a Functional Markup Language, FML [7].

² <http://cadia.ru.is/projects/bmlr/>

In this paper we will not discuss the mechanisms that produce FML automatically – this will be the topic of a future paper – here we will focus on the design of FML. We propose an outline of what FML could contain, and present a starting point for its creation.

Although BML and FML will very likely share ideas, especially related to temporal issues and synchrony, a number of issues will necessarily be different between these two languages. In particular, time and temporal dependencies in FML will certainly be represented in a coarser way than in BML (just as BML represents time more coarse-grained than the final (frame-based) animation level), as “plans” at this level span larger chunks of time and can thus be seen as providing a “rough outline” or specification for the next planning level below. This process of iterative/depth-first³ construction, as proposed by the SAIBA framework, is represented in Figure 2.

A few words are required to clarify background assumptions. First, we look at dialogue as a continuous, realtime process, turntaking being a case in point, which requires dynamically negotiated role acceptance including who has turn, how multimodal “signals” are interpreted and used, how willing the parties are in trying to understand each other, go along with premises put forth, etc. [7]. (Most often such negotiation is implicit and goes unnoticed by dialogue participants.) We also assume agent-orientation: A participant is at any point in time in charge of *only his own end* of such a realtime activity, but of course in all except the most extreme cases has an ability to affect the other party in many ways by his own behavior.

2. WHAT IS A FUNCTION?

In the present context, by “function” we mean *the effect that an action is intended to have* in a particular multimodal communicative interaction, either on the body of the actor him/herself and/or on the mind/body of the interlocutor(s). What is typically expressed here are inner states such as affect or agreement, and those related to management of the interaction itself, including the exchange of turns. The realization of these functions relies on the coordination of a wide range of behaviors including prosody, vocal fillers, head motion, body posture and eye gaze, all of which are specified further at the behavior-level.

3. FML: AN OUTLINE

FML must describe the effect that an intended action or plan should have on the environment, most obviously the agent itself, that must express that function; in line with the SAIBA framework it leaves out, however, morphological considerations that are intended to be composed at runtime by one or more (mostly as-of-yet unspecified) engines/mechanisms. As with BML this is the research part: The language does not specify *how* an

³ It should be noted that SAIBA does not specify that planning needs to proceed in a top-down manner; it is well conceivable that higher levels take proposed BML as input and generate FML as a way to make sure that the behavior to be performed does not work against the agent’s goals at that point in time (example: I really want to scratch my head but my boss has told me not to). The same could be done for producing a hypothesis for what another person’s behavior means at the functional level; in this case the BML is used to describe the surface form of the actions observed; the FML would represent hypotheses for what it is intended to achieve.

FML specification is turned into actual expressed behavior, whether through BML or some other way.

We propose that FML be based on labels which refer to basic functions of multimodal communicative actions/goals, and constraints on those functions, which themselves can be divided into functional categories. An example is the communicative function to express moderate happiness: The function *express_happiness* has a constraint describing its amount, say "medium" or "0.5".

There are several things that need to be taken into account for FML to be successful. First, we must provision for coarse-grained temporal constraints. Second, we need a prioritization scheme so that an FML Engine can be given instructions as to how to solve conflicting functions. Such a scheme could be represented as any other type of constraints on functions. Third, we need to classify functions in to groups that help a human designer use FML; Thórisson's scheme in the Ymir/Gandalf system [8] of splitting them into *Topic Functions* and *Envelope Functions* provides an example of a step in this direction.

3.1 Temporal Constraints

Temporal constraints at the functional level of description tend to be much more coarse-grain than those at the lower levels termed "behavior" and "execution". At the execution level one has to deal with frames and milliseconds; at the intermediate behavior level we deal with temporal relationships of bits of multimodal events such as e.g. gaze, grasps and body stance; at the functional level we specify temporal relationship between what we could call "plan chunks". These chunks refer to parts of a plan that implements the form for e.g. a set of inter-related propositions to be expressed, for instance directions on how to get from one city to another. Each such plan chunk will typically consist of several multimodal acts at the behavioral level. To take an example, in a full plan consisting of several chunks intended to help someone decide where to take a walk in the forest, pointing at a map and saying "You start here [deictic gesture; looking at map], and walk all the way through the forest [tracing with finger], and end up here [finger stops], will take you approximately 1 hour [gazing back at interlocutor]" would be one plan chunk containing (roughly) three BML chunks.

To specify temporal relationships between functional plan chunks we propose to start with simple synchronization primitives such as:

```

must_end_before(a,b,T)
execute_anytime_during(a,b,T)
start_immediately_after(a,b)
start_sometime_after(a,b,T)
start_together(a,b,...z,T)
end_together(a,b,...z,T)

```

These are relatively self-describing; *a* and *b* are plan chunks whose relation is described with the primitive, where *b* is the reference; *T* is an optional parameter that describes a maximum boundary or tolerance which can be provided by the designer or even computed dynamically, based on context.

3.2 Prioritization Scheme

In the Ymir/Gandalf system Thórisson [8] proposed three main levels of prioritization: *Reactive*, *Process Control* and *Topic*, each

one of a lower priority than the prior, respectively. If a reactive behavior is required while a behavior of a different priority is executing, the reactive behavior takes precedence. If a process control-level behavior is requested while a topic-level plan is being executed the latter one will have to yield. This prioritization scheme has been proposed as a cognitive theory of human dialogue organization [7]. More importantly for the present discussion, a key goal of such a prioritization scheme is to enable a designer to stop worrying, to some extent, about unwanted interactions between conflicting behaviors.

Generally speaking, BML maps to a reactive level while FML to the process control level (and partly the Topic level). Compared to these priority levels in Ymir architecture, however, this mapping is not 1:1 because Ymir separated a Behavior Lexicon from perception-driven decision/planning mechanisms while SAIBA proposes a different split, as already described in the Introduction. Nonetheless, the comparison can provide a rough sketch for prioritization scheme in FML.

3.3 Classification

The classification of behavioral functions will aid the designers of multimodal dialog systems at different levels. At the highest level, the designer will see a rough outline of the human communicative capacity of a system by noting what general kinds of function specification are available. At a much lower level, a designer can expect that functions within a certain category will share some specification characteristics (such as types of constraints) or share a representation of common plan chunks or structures (such as turns or participants).

Choosing a classification scheme that embraces all prevailing perspectives on communicative function is not easy, but we have to start somewhere. The research community has more or less come to an agreement about the existence of a category of communicative functions that serve to *coordinate* a multimodal dialog. The functions in this category have been called envelope, interactional or management functions [8][10][11]. Examples gathered from a range of multimodal dialog projects are shown in Table 1a (these tables are replicated from [10]).

Table 1a: ENVELOPE/INTERACTION FUNCTIONS	
Function Category	Example Functions
Initiation / Closing	<i>react, recognize, initiate, salute-distant, salute-close, break-away</i>
Turntaking	<i>take-turn, want-turn, give-turn, hold-turn</i>
Speech-Act	<i>inform, ask, request</i>
Grounding	<i>request-ack, ack, repair, cancel</i>

Another category covers the actual content that gets exchanged during a dialog. Given that the envelope functions are doing a proper job in an ongoing dialogue, topic functions have a better chance of being achieved. Typically this is the deliberate exchange of information, which gets organized and packaged in

information chunks⁴ that facilitate uptake/interpretation in an interlocutor. Another set of function examples have been gathered for Table 1b.

Table 1b: TOPIC/CONTENT FUNCTIONS	
Function Category	Example Functions
Discourse Structure	<i>topic, segment</i>
Rhetorical Structure	<i>elaborate, summarize, clarify, contrast</i>
Information Structure	<i>rheme, theme, given, new</i>
Propositions	<i>any formal notation (e.g. "own(A,B)")</i>

It has been suggested that a useful distinction could be made between functions that carry deliberate intent and those that merely give off behavior involuntarily [5]. Examples of such functions are shown in Table 1c.

Table 1c: MENTAL STATE AND ATTITUDE FUNCTIONS	
Function Category	Example Functions
Emotion	<i>anger, disgust, fear, joy, sadness, surprise</i>
Interpersonal Relation	<i>framing, stance</i>
Cognitive Processes	<i>difficulty to plan or remember</i>

This classification, along with each of the named functions, is a proposal for actual FML tags, which can be discussed further at this workshop.

4. CONCLUSIONS

SAIBA is an important effort in coordinating and advancing research on multimodal behavior generation and the specification of FML is a key element. For FML to be successful, three things in particular have to be taken into account: (1) Temporal constraints at a coarser level of granularity than the BML level; (2) A prioritization scheme that helps arbitrate conflicts and supports reactivity; (3) Classification of FML tags into categories that help human designers make sense of communication capabilities as well as for identifying groups of functions with similar parameterization. These notes should serve as seeds for a discussion that is highly relevant to the kinds of real-time multimodal dialog systems being built at CADIA.

REFERENCES

- [1] Kopp, S., B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, H. Vilhjálmsón: Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. *Proceedings of Intelligent Virtual Agents (IVA '06)*, August 21-23. Also published in Springer Lecture Notes in Computer Science (2006)
- [2] Herman, M. and Albus, J.: REAL-TIME HIERARCHICAL PLANNING FOR MULTIPLE MOBILE ROBOTS. In *Proc. DARPA Knowledge-Based Planning Workshop*, Austin, Texas, December 1987, 22-1 – 22-10 (1987)
- [3] Vilhjálmsón, H., Cantelmo, N., Cassell, J. et al.: The Behavior Markup Language: Recent Developments and Challenges. *Proceedings of Intelligent Virtual Agents (IVA '07)*, Vol. LNAI 4722. Springer (2007) 99-111
- [4] Thiebaut, M., Marshall, A., Marsella, S., and Kallmann, M., SmartBody: Behavior Realization for Embodied Conversational Agents, in *Proceedings of Autonomous Agents and Multi-Agent Systems (AAMAS) (2008)*
- [5] Heylen, D., Kopp, S., Marsella, S., Pelachaud, C., Vilhjálmsón, H.: The Next Step Towards a Functional Markup Language. *Proceedings of Intelligent Virtual Agents (IVA '08)*, Springer (2008)
- [6] Vilhjálmsón, H., & Stacy, M.: Social Performance Framework. *Workshop on Modular Construction of Human-Like Intelligence at the 20th National AAAI Conference on Artificial Intelligence, AAAI (2005)*
- [7] Thórisson, K. R. Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perception to Action. In B. Granström, D. House, I. Karlsson (Eds.), *Multimodality in Language and Speech Systems*, 173-207. Dordrecht, The Netherlands: Kluwer Academic Publishers. (2002)
- [8] Thórisson, K. R. Gandalf: An Embodied Humanoid Capable of Real-Time Multimodal Dialogue with People. *First ACM International Conference on Autonomous Agents*, Marriott Hotel, Marina del Rey, California, February 5-8, (1997) 536-7
- [9] Vilhjálmsón, H. and Thórisson, K.R. A Brief History of Function Representation from Gandalf to SAIBA, in the *proceedings of the 1st Function Markup Language Workshop at AAMAS*, Portugal, June 12-16, (2008)
- [10] Vilhjálmsón, H. Representing Communicative Function and Behavior in Multimodal Communication, A. Esposito et al. (eds.), *Multimodal Signals: Cognitive and Algorithmic Issues, Lecture Notes in Artificial Intelligence*, Vol. 5398. Springer (2009)
- [11] Thórisson, K. R. and Jonsdóttir, G. R.: A Granular Architecture for Dynamic Realtime Dialogue. *Proceedings of Intelligent Virtual Agents (IVA '08)*, Tokyo, Japan, September 1-3 (2008)

⁴ In other work we have used the concept of “thought unit” as the smallest unit that is ready to be turned into the equivalent of a BML-specified behavior [11].