Towards a Common Framework for Multimodal Generation: The Behavior Markup Language

Stefan Kopp[1], Brigitte Krenn[2], Stacy Marsella[4], Andrew N. Marshall[4], Catherine Pelachaud[3], Hannes Pirker[2], Kristinn R. Thórisson[5], and Hannes Vilhjálmsson[4]

[1] Artificial .Intelligence Group, University of Bielefeld, Germany
skopp@techfak.uni-bielefeld.de
[2] Austrian Research Institute for AI (OFAI), Vienna, Austria
{brigitte, hannes}@ofai.at
[3] IUT de Montreuil, University de Paris 8, France
c.pelachaud@iut.univ-paris8.fr
[4] Information Sciences Institute, University of Southern California USA
{marsella, amarshal, hannes}@isi.edu
[5] CADIA, Dept. Of Computer Science, Reykjavik University, Iceland
thorisson@ru.is

Abstract. This paper describes an international effort to unify a multimodal behavior generation framework for Embodied Conversational Agents (ECAs). We propose a three stage model we call SAIBA where the stages represent intent planning, behavior planning and behavior realization. A Function Markup Language (FML), describing intent without referring to physical behavior, mediates between the first two stages and a Behavior Markup Language (BML) describing desired physical realization, mediates between the last two stages. In this paper we will focus on BML. The hope is that this abstraction and modularization will help ECA researchers pool their resources to build more sophisticated virtual humans.

1 Introduction

Human communicative behaviors span a broad set of skills, from natural language generation and production, to coverbal gesture, to eye gaze control and facial expression. People produce such multimodal behavior with ease in real-time in a broad range of circumstances. The simulation of such behaviors with computer-generated characters has, by now, a history of more than ten years [15][1]. A number of approaches have been presented in the field, geared toward specific aspects of generating multimodal behavior, e.g. facial expressions and gesture synthesis. All represent models of a production process in which certain knowledge structures are identified and transformed. Such knowledge structures include representations of communicative intent, lexicons that define available behaviors and their particular overt forms, and rules as to how communicative intent and affective state is mapped onto them.

At the AAMAS 2002 workshop "Embodied conversational agents - let's specify and evaluate them!" it became obvious that most researchers were building their own

J. Gratch et al. (Eds.): IVA 2006, LNAI 4133, pp. 205 – 217, 2006.

behavior and functional languages. While diversity is important, another "Gesticon" workshop in 2003 made it clear that a lot of similarities existed among the approaches. To avoid replication of work, as well as to allow for sharing modules, a push was initiated to develop a common specification. In April 2005, a group of researchers in the area of multimodal communication and computer animation came together at Reykjavik University to further the integration and development of multimodal generation skills for artificial humans [18]. Our goals were (1) to frame the problem of multimodal generation in a way that allows us to put it into computational models; (2) to define planning stages of multimodal generation and to identify the knowledge structures that mediate between them; (3) to render these stages and knowledge structures into a framework that lays down modules and interfaces, enabling people to better work together and to use each other's work, that has been directed to different aspects of multimodal behavior, with a minimal amount of custom work. In previous efforts we started by clarifying terminologies such as representation vs. markup vs. scripting languages [9].

In this paper we describe our latest results in this ongoing process. In Section 2, we begin by looking into four existing languages: BEAT, MURML, APML and RRL. Our goal is to bring together our experiences with these languages and to derive a powerful, unifying model of representations for multimodal generation. We present such a model, the SAIBA framework, in Section 3. Two important representation languages emerged as part of this framework. These languages are meant to be application independent, graphics model independent, and to present a clear-cut separation between information types (function versus behavior specification). We will go into one of those languages, the Behavior Markup Language (BML), in more detail in Section 4, and then conclude with remarks on the next steps.

2 Prior Approaches

A number of researchers have construed representation languages for capturing the knowledge structures that were identified as involved in the generation of multimodal behavior. We start here by analyzing four broadly used languages, all being XML compliant. While there are certainly more languages being employed out there (e.g. MPML; [12]), the languages considered here provide a good overview of previous approaches, and allow us to compare the assumptions that underlie their generation models.

One principal commonality among these and related previous systems is the separation of content- and process-related processing. For example, the Ymir architecture used to implement the Gandalf humanoid clearly separated dialog planning and social interaction control [16][17]. The argument behind this was that what an agent chooses to say in a given situation is highly domain-specific, whereas the ability to deliver that content through social interaction is a broad re-usable skill. Consequently, verbal responses related to dialog topic (content) were generated by a separate process, based on the user's interpreted communicative act (the multimodal version of a speech act), using an abstract frame-based representation. The surface form, however, of this content and all necessary process-related responses (turntaking signals, gaze, head movements, gesture, paraverbals), was generated by a realtime, rule-based planner (called Action Scheduler) in incremental chunks of 200-1200 msec