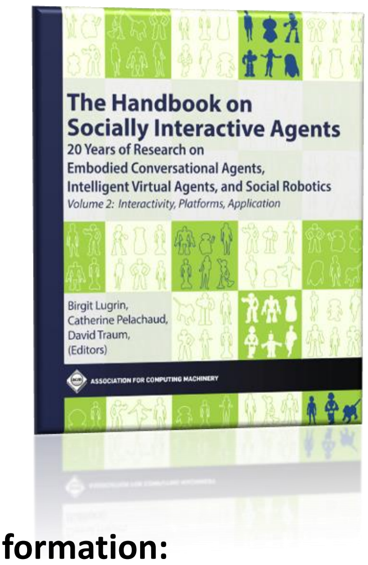
The Fabric of Socially Interactive Agents: Multimodal Interaction Architectures

Stefan Kopp and Teena Hassan



The Handbook on Socially Interactive Agents

20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics

Volume 2: Interactivity, Platforms, Application

Birgit Lugrin,
Catherine Pelachaud,
David Traum,
(Editors)

ASSOCIATION FOR COMPUTING MACHINERY

Citation information:

S. Kopp and T. Hassan (2022). The Fabric of Socially Interactive Agents: Multimodal Interaction Architectures. In B. Lugrin, C. Pelachaud, D. Traum (Eds.), *The Handbook on Socially Interactive Agents – 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics*, Volume 2: Interactivity, Platforms, Application (pp. 77-111). ACM.

DOI of the final chapter: 10.1145/3563659.3563664

DOI of volume 2 of the handbook: 10.1145/3563659

Correspondence concerning this article should be addressed to

Stefan Kopp skopp@techfak.uni-bielefeld.de and Teena Hassan thassan@uni-bremen.de

16 The Fabric of Socially Interactive Agents: Multimodal Interaction Architectures

Stefan Kopp and Teena Hassan

16.1 Motivation

The field of socially interactive agents has emerged out of many different approaches to create technical systems that can engage in natural human-like conversation, nonverbal communication, multimodal dialog or emotionally aware interaction. Many of these directions have grown into research fields in their own right (e.g. Social Signal Processing, Affective Computing, Social Robotics, Natural Language Processing, Spoken Dialog Systems/Conversational Agents), with specific foci, methodological approaches, and technologies. In result, the variety but also the specialization of approaches has been growing. For example, we now have elaborated methods for the recognition or synthesis of social signals in specific modalities (e.g. speech, prosody, facial expressions, gaze, or gesture), their fusion to extract semantic or pragmatic meaning, or the planning of interactive behavior to fulfill emotional or relational goals. This gives rise to an *integration problem* as, in social face-to-face interaction, many of these abilities and skills need to be at work at the same time and in an integrated and coordinated manner [Gratch et al. 2002]. A crucial question is thus not only how the various capabilities and features of a Socially Interactive Agent (SIA) can be realized technically, but also how they can come to play together within a functionally complete interactive virtual agent or social robot.

The present chapter focuses on concepts and methods to realize and integrate approaches to achieve abilities needed for conversational *multimodal interaction*. We will thereby go beyond the usually separate perspectives towards (and solutions for) processing multimodal input or generating multimodal output. Instead we aim to provide an overview of how such techniques are mapped out and integrated in current virtual agents or social robots by means of a suitable *interaction architecture*. From a practical point of view, an agent architecture may be regarded as a collection of specialized modules linked together by means of inter-process communication. From a conceptual point of view, however, it has to answer the question of how the underlying complex computations that are required to act like a socially intelligent agent in real-time interaction can be organized and orchestrated. It thus provides the underlying structure that provides the constraints and affordances both for how single modules are to operate, as well as for how the agent as a whole is able to (inter-)act in a consistent, timely and believable manner, and how it will thus appear as an interaction partner.

Knowing about the challenges, principles and approaches for developing multimodal interaction architectures of SIAs has become increasingly important for researchers and practitioners alike. Interacting with today's social robots or virtual agents is often characterized by stereotypical behavior or slow response times, resulting in unnatural clumsiness or disfluencies. Human conversational interaction, in contrast, is characterized by an inherent multimodality and high responsiveness with which cooperative interactants construct their contributions. For example, even while producing communicative actions, speakers attend to and elicit reactions from their addressee [Clark and Krych 2004]. Depending on this immediate feedback, speakers can re-plan the remaining part(s) of their communicative act, adapt it to the addressees' needs, put it on hold, interject a sub-dialog, and continue at the point of interruption. All of this is done in such an effortless, smoothly coordinated and seemingly natural way that it is not even apparent that difficulties were payed attention to or that plans were changed mid-way. Thus, acting in a conversation is not solely based on extensive planning ahead and deep representational models. Instead, interaction partners, while being guided by overall goals and strategies, are also highly sensitive to the partner's verbal and nonverbal behavior and are able to alter their multimodal utterances accordingly. These abilities are crucial for human-like fluent conversation – and they imply important demands for how to construct interactive agents at the architectural level.

In the following, we will start by identifying overarching requirements and criteria for multimodal interaction architectures. We then review different approaches and concepts that have been put forward in the fields of (embodied) conversational agents and social robotics. Finally, we will point out main challenges and directions to be pursued in order to succeed in weaving the fabric of truly socially interactive and intelligent agents.

16.1.1 Requirements for multimodal interaction

Architectures of social agents or robots are usually designed with a particular functional goal in mind, such as joint attention, empathy, imitation, or interactive learning [Breazeal et al. 2004, Duffy et al. 2005]. In this chapter we discuss how the specific components and layout of SIA architectures enable (or hamper) multimodal conversational interaction with a human user. We start by identifying a number of requirements that an SIA needs to meet in order to provide multimodal interactivity to its user.

One obvious requirement is the ability to *recognize* the relevant verbal and nonverbal input as well as to *generate* convincing multimodal output. A main distinguishing aspect is thus the number and kinds of *modalities* supported when interacting with the agent. Most virtual interactive agents and social robots have included visual and auditory sensory modalities (e.g. [Baxter et al. 2013, Dodd and Gutierrez 2005, Kasap and Magnenat-Thalmann 2010, Kędzierski et al. 2013, Matsuyama et al. 2016]). In addition, few virtual agents [Bosse et al. 2018] but several social robots (e.g. Breazeal et al., 2003; Pepper [SoftBank-Robotics 2021b], Paro [Shibata 2012]) support tactile stimuli to perceive or produce touch. However, a modality goes beyond a mere sensory channel and must be considered a *semiotic system* that affords certain semantic and pragmatic functions by means of specific displays or signals conveyed over a certain sensory channel. For example, spoken language as a vocal modality allows for conveying symbolic content, intonation can add "analog" acoustic cues of prominence or stance, and gesture as a visual modality lends itself to communicate indexical or iconic meaning. Artificial agents may even add other non-human modalities to this. Multimodal communication, then, arises from the combination and integration of those different ways of communicating meaning. It thus involves not only the processing and generation of single verbal and nonverbal behaviors, but also their interpretation and embedding in *coherent multimodal ensembles* whose parts are coordinated in form, meaning, or pragmatic function as well as in their temporal arrangement. The corresponding multimodal coherence and cross-modal relations are vital for a recipient to be able to resolve the overall intended meaning.

A second requirement, pointed out by Cassell and colleagues [Cassell et al. 2000] in their seminal work on embodied conversational agent frameworks, is to be able to deal with behavior in terms of its multiple conversational *functions* (e.g. conveying content, representing socially, managing conversation) and based on an understanding of a dialog state that can involve multiple threads of communication. This relates to the grounding of multimodal behavior processing into models and representations of (changes of) an interaction state and the selection of multimodal behavior in order to change this state according to interaction goals or policies. A related requirement is that multimodal behaviors need to support a sufficient degree of *expressiveness* that is needed for the communicative demands and believability of the human-agent interaction at hand.

Thirdly, multimodal interactions unfold at *multiple timescales*, from milliseconds between eye-contact and a head nod, to longer periods of time for utterances or even larger discourse segments [Cassell et al. 2000]. Across these timescales, multimodal conversational behavior must be sufficiently *fluent and continuous*. Unwanted and unnatural lags, hesitations, or disfluencies can lead to interaction problems (e.g. overlapping speech with dialog systems) as well as ambiguous, incoherent meaning (e.g. when pointing to an object too late). Multimodal SIAs thus need to be able to manage multimodal behavior in realtime, at multiple timescales in parallel and with the corresponding fluency. Further, it is often emphasized [Kopp et al. 2014, Schlangen and Skantze 2011] that fluid conversation hinges upon fast and reciprocal adaptation between the interlocutors. For example, in a multimodal interaction one often has to adapt one's own behavior to the interlocutor's actions in an online and well-timed fashion, e.g. to keep or take the floor [Levinson and Torreira 2015], to respond to communicative feedback and interruptions, or to entrain and align with one another [Lakin et al. 2003]. Consequently, another requirement for SIAs is fast *responsiveness*, *adaptability*, and *interruptability* in their multimodal behavior.

16.2 Models and Approaches

A large variety of SIAs have been designed to support some form of multimodal interaction. In this vein, different architectural approaches to organize and realize the processing of (multi-) modal input and output have been employed in IVA or SR. Yet, existing systems fulfill the above-mentioned behavioral requirements only to a partial and different degree. In the following, we will discuss the architectures that have been developed in relevant fields. They can be compared and assessed with respect to a number of features:

- **Modalities**: What is the number and kinds of modalities included, and what is the degree of multimodal integration (fusion/fission)?
- **Methods/components**: What are the techniques and approaches used for recognition/interpretation, generation, and planning of multimodal behavior (at the task level as well as the social-relational level)?
- **Processing structure**: How is control and processing organized across different routes (pathways, streams) or at different levels (deliberative, reactive, associative)? What is the general way of processing input/output over time (sequential/parallel, chunked/incremental)?
- **Interactive adaptivity**: Is the social interaction dynamics with its reciprocal feedback loops taken into account? How are processing and generation connected to support fast adaptivity? What kind of cross-modal interactions are considered?
- **Technical applicability**: Is the approach specific to virtual or robotic agents? How modular, inter-operable and portable is the approach?

In order to provide a systematic overview and to make approaches comparable, we will characterize them according to which parts of a conceptually "complete model" of multimodal interaction they support. A schematic of this conceptual model is shown in Fig. 16.1. It comprises three columns for (1) processing multimodal input, (2) mapping responses, and (3) generating multimodal output. Each column, in turn, comprises different levels of processing, from sensory-motor behavior to high-level conversational and socio-relational functions. For the input column (left-hand side), this relates to the common processing pipeline from sensing data, to recognizing features or patterns, to interpreting them with regard to meaning or interactional functions. For the output column (right-hand side), the stages correspond to a standard generation pipeline [Kopp et al. 2006] that involves determining modalities and behavioral forms (e.g. words, intonation, gestures, expressions) to fulfill a given intent, turning them into actual synchronized behavior with the bodily resources of the agent, and finally acting them out overtly. The middle column maps between input and output at different

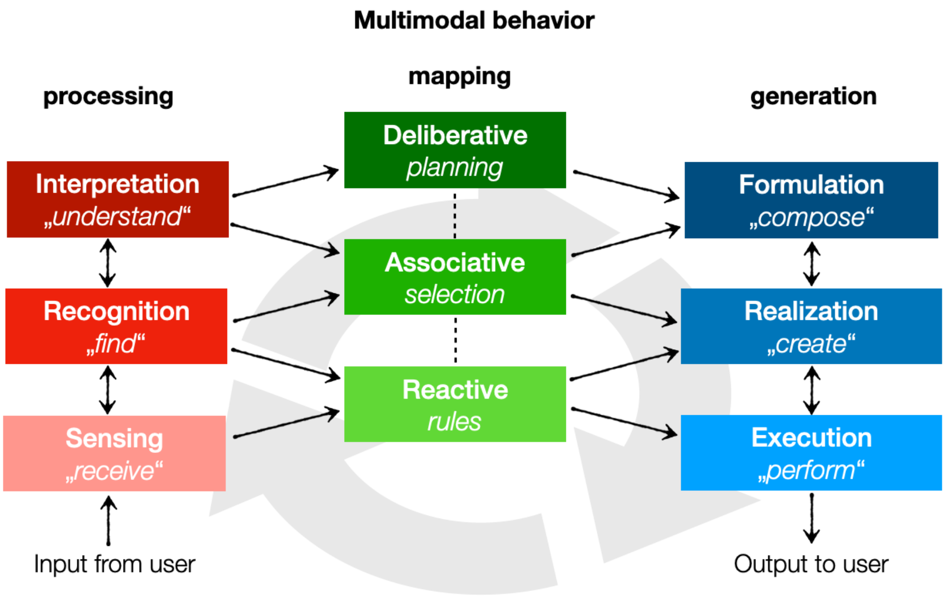


Figure 16.1 Schematic of different processes and pathways in multimodal interaction architectures.

levels of decision-making processes, from reactive (based on hard-wired rules), to associative (selecting from a given set of alternatives), to deliberative (planning possibly new responses).

It is important to note that different pathways of processing are possible through the columns and layers (and are actually taken in existing systems). For example, multimodal signals can be processed up to interpreting a user state that is then mapped via associations to predefined outputs (circumventing planning of goals, content, or forms). Note also that each of these boxes can be more or less modality-specific or multimodal. For example, recognition can work on different modalities separately with specific models whose outputs are combined afterwards (so-called “late fusion”), or it can work on multimodal data (after an “early fusion”) to find larger, integrated patterns or features.

In the following, we will discuss different multimodal interaction architectures that have been applied in SR and IVA. We will thereby characterize them and make them directly comparable by mapping their architectural components to the schematic shown in Fig. 16.1, using the same color code to relate specific parts of the architecture.

16.2.1 Embodied conversational or virtual agent architectures

Virtual agents or embodied conversational agents (ECA) are graphically rendered characters designed to support a human-like conversational interaction with a human user [Cassell 2001].

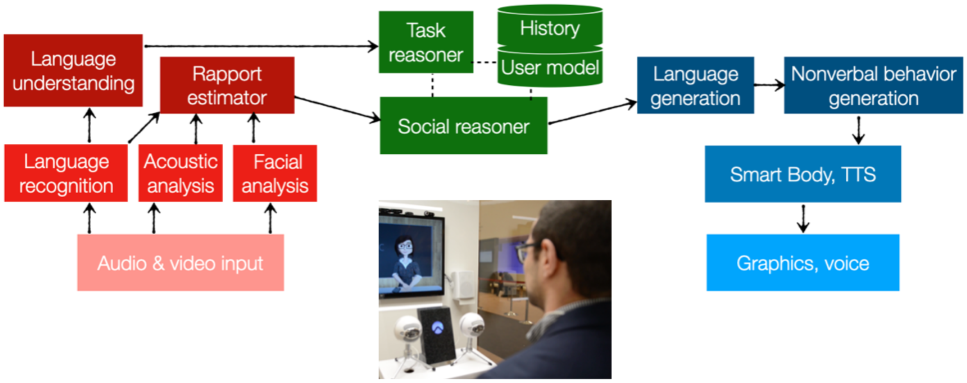


Figure 16.2 Example of a single-route architecture for plan-based multimodal behavior: Socially-Aware Robot Assistant, abbreviated SARA (based on [Matsuyama et al. 2016]). Inset shows a person interacting with SARA (© 2021 Justine Cassell).

A large number of such agents have been developed, focusing on different kinds of socio-communicative behavior, abilities, or application scenarios. Throughout this endeavor, a range of architectural principles and models have emerged.

Single-route architectures Many ECA systems have focused on producing socially appropriate multimodal behavior to achieve, e.g., engagement, rapport, trust, or empathy. The architectural layout consists of a single route with multiple consecutive processing steps, usually involving high-level state representations and planning-based behavior generation. For example, so-called "relational agents" employ specific planners to increase and maintain rapport with the user [Papangelis et al. 2014] or to achieve long-term engagement [Bickmore et al. 2010]. One example is the proposed SARA architecture [Matsuyama et al. 2016] (Fig. 16.2), in which a task planner and a social reasoning component are combined with a memory-based model of the user and previous interactions, as well as with corresponding modules for social behavior interpretation (here, estimating rapport from utterances, conversational strategies, acoustic features, and 3D facial landmarks) or behavior generation (natural language generation, nonverbal behavior generation). Together, these modules form a single *deliberative route* of multimodal processing along which the agent produces socially attuned behavioral responses to complete user inputs. The focus in these systems lies on the functional quality of the produced behavior, and less on its embedding in a fluent and dynamic conversational interaction.

Dual-route architectures Complete ECA systems aim to enable an efficient and robust face-to-face conversational interaction (e.g. Rea [Cassell 2001], Max [Leßmann et al. 2008], or Greta [Bevacqua et al. 2010]). These systems employ a dual-route architectural layout (see

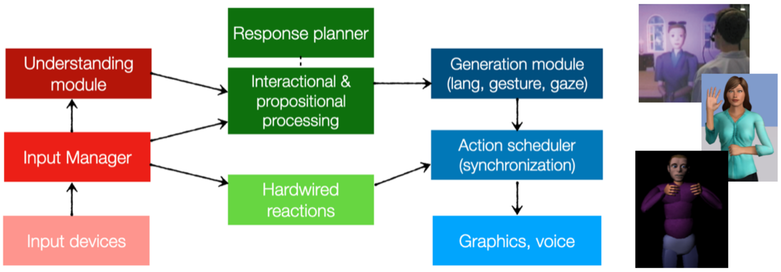

Response planner
Understanding module
Interactional & propositional processing
Generation module (lang, gesture, gaze)
Input Manager
Action scheduler (synchronization)
Hardwired reactions
Input devices
Graphics, voice

Figure 16.3 General structure of a dual-route architecture with parallel deliberative and reactive processing of multimodal behavior (Inset shows three conversational agents based on this layout: REA [Cassell et al. 2000]; Greta [Bevacqua et al. 2010] (© 2021 Catherine Pelachaud); Max [Leßmann et al. 2008] (© 2021 Stefan Kopp)).

Fig. 16.3). As described above, the *deliberative* route comprises higher-level processes for reasoning and planning of desired interactive functions and behaviors. This is usually based on classical natural language processing pipelines in Spoken Dialog Systems, which include some form of semantic decoding and dialog state tracking, based on which the system output is determined through some form of (pre-)planned policy. ECAs employ a similar pipeline but extend it to interpreting and planning multimodal communicative behavior (e.g. gaze, gesture, body posture, facial expression) for conversational or socio-relational functions (e.g. dialog grounding, turn-taking, attention, politeness, empathy). The underlying models are based on dedicated representational and decision-making models, either symbolic and rule-based (classically) or implicit and learned from data (more recently). A second *reactive* route, in contrast, implements more direct mappings from perceptual events to overt behavior. This route is required to support fast social feedback loops, e.g., in continuous gaze-tracking or behavioral mimicry (by mapping user location or user movement to animated adjustments of the agent body). It is also necessary for an agent's subtle and dynamic expressiveness, e.g., through emotional facial expressions (by mapping internal affective states to animated facial features). Both routes rest on a behavior realization mechanism that is in charge of arbitrating, combining, synchronizing, and finally producing the eventual output behaviors.

Multi-directional, incremental architectures Many of the existing ECAs build on the dual-route architecture layout, characterized by concurrent processing along a deliberative and a reactive route. However, each route itself realises a sequential processing of input/output units at a corresponding level of abstraction and granularity. The growing awareness of the role of multimodal behavior in the dynamic grounding of dialog (see Sect. 16.1.1), however, and the view that ECAs ultimately need to be able to support these mechanisms for collaboratively co-constructing mutual understanding [Kopp and Krämer 2021], has led to further advancements

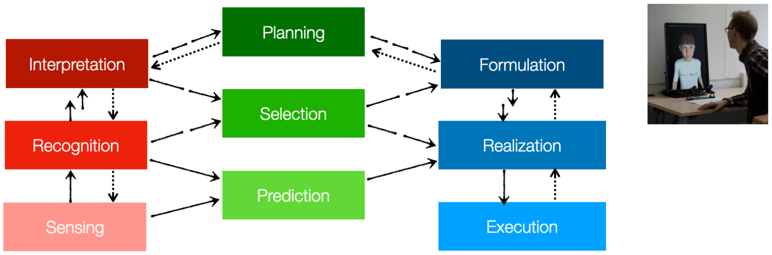


Figure 16.4 Architectural layout for fluid, real-time conversational interaction with multi-directional flow of information, incremental processing, and prediction-based behavior (Inset: Conversational agent ”Billie” [Kopp et al. 2018]; © 2021 Hendrik Buschmeier).

at the architectural level. They can be summarized under two key concepts: *Multi-directional flow of information* and *incremental processing*.

Incremental processing has been frequently identified as a key principle for natural dialog modeling with phenomena such as fluent turn transitions, interruptions, disfluencies, or fast adaptations to the interlocutor. For example, the ‘How Was Your Day?’ prototype [Crook et al. 2012] for coping with barge-ins employed a ‘long’ loop for intent planning and a shorter loop to handle interruptions, back-channel feedback and emotional mirroring. However, the authors note that the use of *incrementality* would have made their design more elegant and efficient. Schlangen and Skantze [Schlangen and Skantze 2011] described incremental dialog agents in terms of abstract modules that communicate via incremental units (IUs) that are extended in a step-wise fashion before being finally committed. Several implementations of this model have been developed and many aspects of language-based dialog have been successfully modeled within this incremental processing framework (e.g., speech recognition, natural language understanding [Atterer et al. 2009], dialog management [Traum et al. 2012], natural language generation [Skantze and Hjalmarsson 2013], speech synthesis [Buschmeier et al. 2012]). A few recent approaches have tried to apply this principle to process and integrate multimodal input, e.g., speech and gesture [Han et al. 2018].

The second extension refers to enabling a multi-directional flow of information. This includes the passing of feedback information from downstream components back to higher-level modules. The SAIBA (Situation, Agent, Intention, Behavior, Animation) framework for multimodal generation [Kopp et al. 2006, Vilhjálmsson et al. 2007] provided markup languages for specifying mltimodal behavior (BML) and its functions (FML) to be processed by subsequent modules. Additionally, it emphasized the importance of feedback to inform planning modules about the extent to which their decisions were being realized. Further, a multi-directional flow of information also refers to information flowing internally from generation to processing (i.e., from right to left in our schema). It has often been stressed that in the brain top-down information helps to bias or prime sensory processing towards specific information or to resolve ambiguities based on contextual information (cf. [Teufel and Nanay 2017]). Yet, very few SIA architectures have modeled input processing components to receive information from higher-level components of the architecture. Nijholt et al. [Nijholt et al. 2008] proposed a first approach to directly link the timing of an agent's behavior to the *predicted timing* of interlocutor events. This enabled a finer degree of temporal coordination with the user's motion, as demonstrated in a dancer agent, a virtual orchestra conductor and a virtual fitness trainer. The Artificial Social Agent Platform (ASAP) [Kopp et al. 2014, Van Welbergen et al. 2014] proposed extensions to the behavior markup language (BML) in order to bidirectionally link sensory input processing and generation of agent's behavior.

Overall, ECA designed for fluid interaction resolve the strict layout of sequential processing architectures, in favor of a more flexible distribution of processing, both with respect to the flow of information as well its temporal organization. As illustrated in Fig. 16.4, incremental processing is applied in particular at higher levels of the architecture, where units of processing (e.g. a full dialog act) are more abstract, arise at a relatively lower rate, and have a larger temporal scope in the underlying overt behavior.

Behavior generation sub-architectures Work in the field of ECAs or virtual agents has traditionally focused on the generation of expressive, communicative multimodal behavior. Consequently, a lot of systems and models have been developed to embody the right-hand side of our architectural schema. Two main approaches have emerged that can be distinguished according to what they start out from. On the one hand, classical approaches to multimodal behavior generation take some form of *communicative intent* as input and map it to multimodal ensembles out of several, mutually coordinated behaviors. On the other hand, an abundance of recent work has approached the problem of generating multimodal behaviors by starting out from some already given behavior and augmenting it with additional behaviors in other modalities. Such a *cross-modal mapping* approach is, for the most part, driven by speech as input modality. We will discuss both approaches in the following, also noting how they have been combined.

Intent-based multimodal behavior generation is generally conceived to comprise a number of processing steps, similar to the output generation branch of dialog systems. This view has been formalized within the SAIBA framework to encompass three main stages, corresponding to (1) intent planning, (2) behavior planning, and (3) behavior realization [Kopp et al. 2006], along with two XML-based specification languages as interfaces between them (Function Markup Language FML and Behavior Markup Language BML). Generally, these stages relate to the subsequent decision steps of determining what to communicate, with which behavioral forms, and finally how to do it overtly. Intent-based generation, for example, may start from a speech act representation plus some emotional state or socio-relational goals (stated in FML). Behavior planning then usually involves natural language generation (NLG) along with the

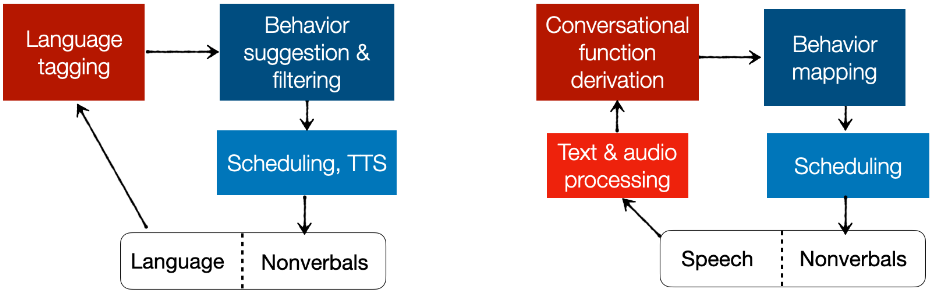


Figure 16.5 Examples of sub-architectures for cross-modal, speech-driven behavior generation: (left) BEAT (based on [Cassell et al. 2004]); (right) Cerebella (based on [Lhommet et al. 2015]).

composition/selection and coordination of appropriate gestures, gaze, or facial expressions. This problem has been tackled using rule/lexicon-based, planning-based, or data/learning-based approaches (see also Chapter "Multimodal Behavior Modelling for Socially Interactive Agents" [Pelachaud et al. 2021] of volume 1 of this handbook [Lugrin et al. 2021]; cf. [Kopp 2013] as well as Chapter 8 of this handbook for an overview), depending on the requirements and criteria for the targeted behavior (e.g. realism, expressiveness, design effort, real-time capability, cognitive plausibility). Behavior realization is then in charge of mapping these multimodal behaviors (usually specified in BML) onto temporally synchronized and coherent articulations or movements. Much work has focused on developing BML-compliant realizers for virtual characters (e.g., ACE [Kopp and Wachsmuth 2004], Greta [Bevacqua et al. 2010], ASAP [Van Welbergen et al. 2014], Smartbody [Thiebaux et al. 2008], or Embr [Heloir and Kipp 2010]), and several approaches were also extended to more flexible timing and motion planning for physical robots [Holroyd and Rich 2012, Salem et al. 2012].

Cross-modal behavior generation recently has become popular due to the direct applicability of Machine Learning methods to large available datasets on human multimodal behavior. The predominant approach is to generate nonverbal behaviors (e.g. gestures, head movements, facial expressions) for a given text or speech output, and the key question is what features are necessary to map from the verbal modality to others in this way. Early speech-driven generation systems applied analysis steps (akin to processing user input) to the linguistic output in order to determine, e.g., speech semantics, information structure, discourse relations or emotions. This additional information is then used to select appropriate behaviors by means of empirically grounded but manually defined rules, or data-based mappings [Cassell et al. 2004, Lhommet et al. 2015] (see Fig. 16.5; see also Chapter 7 "Gesture Generation" [Saund and Marsella 2021] of volume 1 on this handbook [Lugrin et al. 2021]). A key challenge here is that higher-level semantic or pragmatic aspects are necessary, especially for generating

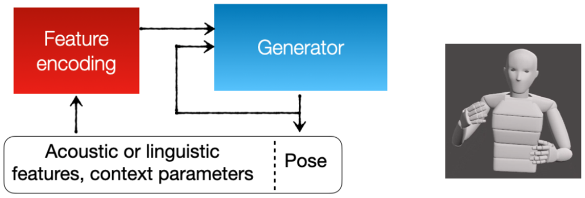


Figure 16.6 Typical architecture of a learning-based model for speech-driven gesture generation (Inset: animated avatar of the "Gesticulator" model [Kucherenko et al. 2020]).

coherent and communicatively meaningful nonverbal behavior like representational gestures that, e.g., depict visual aspects of an object linguistically referred to. These aspects, however, are hard to determine or infer from a semiotically different linguistic input.

Other speech-accompanying behaviors such as small head movements, eyebrow raises or beat gestures, which are less explicitly communicative but nevertheless instrumental for creating a lifelike impression, have been successfully synthesized by mapping directly from the acoustic or verbal features of the speech input. One focus is the speech-driven generation of gesture, for which statistical or, more recently, deep neural network-based models are applied that have been trained to create some form of encoding of the input features and to map it to body postures and movements by way of some generators (usually in an auto-regressive way, i.e. each pose based on the previous one); see Fig. 16.6. Note that these models work in an end-to-end fashion. That is, they comprise both behavior planning and realization and they have been used to drive virtual characters as well as humanoid robots. The techniques that have been explored include probabilistic models [Chiu and Marsella 2011, Ishi et al. 2018], bi-directional long short-term memory networks [Hasegawa et al. 2018], generative models [Kucherenko et al. 2020], mixture models [Ahuja et al. 2020], or generative adversarial networks [Yoon et al. 2020]. Recent approaches have succeeded in producing considerably natural and consistent multimodal behavior, with current work starting to explore how more general contextual parameters such as speaker identity or style can help to further increase output quality (see also Chapter 8 "Multimodal Behavior Modelling for Socially Interactive Agents" [Pelachaud et al. 2021] of volume 1 of this handbook [Lugrin et al. 2021]).

16.2.2 Social robot architectures

Many architectures, more or less cognitively motivated, have been developed for and implemented in social robots (e.g. [Adam et al. 2016, Baxter et al. 2013, Bono et al. 2020, Breazeal et al. 2004, Chao and Thomaz 2013, Laird et al. 2012, Moulin-Frier et al. 2018, Trafton et al. 2013]). Interaction between humans and mobile robots involves several complex challenges that are often absent in the interaction between humans and virtual agents. Given the hardware constraints, the unpredictability of physical actions and their outcomes in the real-world, and the perception challenges posed by the uncontrolled environment, the architectures developed and tested for virtual agents cannot transfer well to robots, especially to mobile robots that co-inhabit or collaborate with humans in the physical world. In order to deal with the complex, high-dimensional, and dynamic domain of human-robot interaction, novel mechanisms for robust perception-action feedback and flexible handling of contingencies arising from failed actions or delayed or failed perception are required, in addition to the integration of safety and privacy-enhancing measures.

In this section, we briefly discuss some of the key interaction architectures that have been developed and tested on social robots. Their design and the modules they are composed of depend on the specific interaction goals being pursued. We start by discussing the sensing and action modalities considered in these works, the interaction goals pursued by them, and the architectural components designed to realize these goals. Then we will turn, again, to the architectural layouts employed in social robots. Depending on the number and type of parallel routes or pathways supported, these can be categorized into single-route, dual-route and multi-directional architectures. We will provide examples for each type of architecture and use the schematic shown in Figure 16.1 to highlight the design principles underlying these architectures.

Sensing modalities Social robots are seldom equipped with only uni-modal sensors for perceiving the external environment. While laser scanners and ultrasound or infrared sensors are used to help the robot navigate safely in the physical environment, sensors for vision, audio and touch are used to perceive information that are especially relevant to initiate and monitor social interaction between a human and the robot. The visual modality is usually used to detect objects and persons in the environment and analyze their properties. For example, Breazeal et al. [Breazeal et al. 2004] used cameras mounted in the eyes of the Leonardo robot to detect and track the face and facial features of the human interaction partner. Malfaz et al. [Malfaz et al. 2011] used the robot Maggie's camera to detect whether there were people standing near the robot. Tanevska et al. [Tanevska et al. 2019] used the eye cameras of iCub robot to detect the face and recognize the facial expressions of the human interacting with the robot. Occasionally, sensors mounted in the environment (external to the robot) are used to augment the visual capabilities of the robot. Breazeal et al. [Breazeal et al. 2004] used cameras fitted on walls behind the robot and above the workspace in order to detect and track objects and people in the broader environment of the robot. More specifically, the overhead camera was used to estimate the head pose and recognize the pointing gestures made by the human interaction partner, as well as to detect and track the state of the shared interaction objects (electric bulbs).

Audio sensors (microphones) are generally used to obtain speech input for automatic speech recognition in order to understand a limited vocabulary of commands or instructions given by the humans (cf. [Breazeal et al. 2004, Malfaz et al. 2011]) and to subsequently extract pre-defined human communicative intents based on simple rules applied to the text recognized from speech (cf. [Adam et al. 2016]). In contrast to vision and audio modalities, touch allows humans to interact with the robot through direct physical contact. Tactile sensors are attached to different parts of a social robot's body (e.g. head, shoulders, chest, back, abdomen, etc.) and they can be used to perceive different properties of human touch, e.g. size of the area touched, the pressure of the touch, the duration of the touch, etc. Touch is an important modality for triggering reactive movements or sounds reflecting liveliness, irrespective of whether the social robot is a humanoid, is animal-like, or has an abstract or hybrid form.

Vision, audio and touch are mostly used separately and serve different perception goals. Even when they share the same goal, they are often combined in a logical-OR fashion. For example, Malfaz et al. [Malfaz et al. 2011] used face detection and speech recognition as redundant channels to determine whether the robot is surrounded by people or is alone. Tanevska et al. [Tanevska et al. 2019] regulated how comfortable iCub felt during social interactions depending on whether and how often it saw a face or felt a touch. A fusion of multiple modalities to infer the state of a human or an object in the environment is not common, which could be partly due to the challenges involved in synchronizing and matching the information provided by the different modalities. Perception outputs (or, percepts) are usually combined with inference rules stored in memory to construct beliefs about the self, the interaction partner and the objects in the environment, and also to determine the internal states of the robot, e.g. emotions (cf. [Lisetti and Marpaung 2007]) or comfort level (cf. [Tanevska et al. 2019]). These beliefs and internal states influence the deliberate behaviors generated by the robot.

Action modalities The utility of social robots derives from their ability to act or behave socially in physical environments with human presence. The social behaviors that they can exhibit depend primarily on the action modalities that it possesses. This in turn depends on the mechatronic design of the robot. In the case of humans, speech, facial expressions, head and hand gestures, torso movements, and locomotion constitute the key overt modalities for expressing social behavior. Humanoid social robots (e.g. Zeno [Hanson-Robotics 2007], iCub [Metta et al. 2008]) possess several or all of the human social behavior modalities, however with reduced degrees of freedom (i.e. fewer movable joints), limited ranges of motion, or alternate types of movements (e.g. locomotion by rolling instead of walking). Social robots with animal-like appearance (e.g. Leonardo [Breazeal et al. 2004], Miro [Ltd. 2020], Paro [Shibata 2012]) support further action modalities, e.g. the movement of ears or tail. In addition, social robots, especially non-humanoid robots having an abstract or cartoon-like appearance (e.g. Pepper [SoftBank-Robotics 2021b]), often possess artificial modalities based on e.g. color LEDs or display screens. While most social robots are mobile, there are also stationary social robots, e.g. the Furhat Robot or the robot reeti® [Robopec 2021], which serve mainly as communicative and expressive robots. The Furhat robot has a 3D facial form on to which a human-like face is projected and virtually animated to give an impression of liveliness. In contrast, the robot reeti® has a head with movable components like eyes, ears, cheeks, and mouth, all of which can be used to show expressions mechanically. Unlike social communication, manipulation is a functionality found less frequently in social robots. Hands and arms, if available, are mainly used for gesturing during single-turn or multi-turn social interactions. However, there can be social interaction scenarios that require social robots to pick, place, carry, or hand-over objects. This would require the robot to integrate task planning and execution in a socially appropriate fashion.

Even when the mechatronic design provides multiple expression modalities, the capability of a social robot to use these modalities to realize multimodal social behaviors depends on the behavior generation and control mechanisms that it is programmed with. While most architectures mention multimodal behaviors, the aspect of temporal alignment of actions across different modalities is often not dealt with explicitly. Even though some works (e.g. [Huang and Mutlu 2014], [Yoon et al. 2019]) have created data-driven models to automatically generate gestures that should accompany speech by learning from annotated human data, these do not include a monitoring component that dynamically adapts the gestures based on run-time synchronization issues. Actions involving the movement of multiple joints are usually defined as a single trajectory in a multidimensional joint space. Although this implicitly describes the temporal alignment of different joints, it makes it difficult to dynamically adapt the motion of individual joints, either due to physical errors or due to a need to merge a new behavior with an ongoing behavior. Such dynamic and seamless adaptation of individual modalities is crucial for generating fluent and naturalistic robot behaviors. However, the ability to incrementally generate and dynamically adapt multimodal behavior still remains elusive to social robots. Having said that, the desired complexity of such behavior generation and control algorithms would depend on the type of the embodiment used (e.g. animal-like versus humanoid) as well as the chosen application domain (e.g. therapeutic versus entertainment).

Single-route architectures The single-route architectures for social robots mostly involve only an *associative* route, where sensory data is processed hierarchically to derive values for internal variables which are then used to select the robot's behavior from a small set of pre-defined actions. For example, Tanevska et al. [Tanevska et al. 2019] used a single-route architecture (see Fig. 16.7) to enable the social robot iCub to adapt its internal drives over time to the specific user it is interacting with. In order to enable this, they included modules to *evaluate* and *adapt* the dynamic "comfort level" of the robot based on the presence or absence of multimodal interaction stimuli, namely face and touch. Based on the current comfort level, the robot decided whether to engage or disengage with the user and accordingly selected the actions to be performed (*behavior selection*). The actions involved the movement of different joints on the head/neck, arms or torso of the robot, and the motion trajectory was adapted

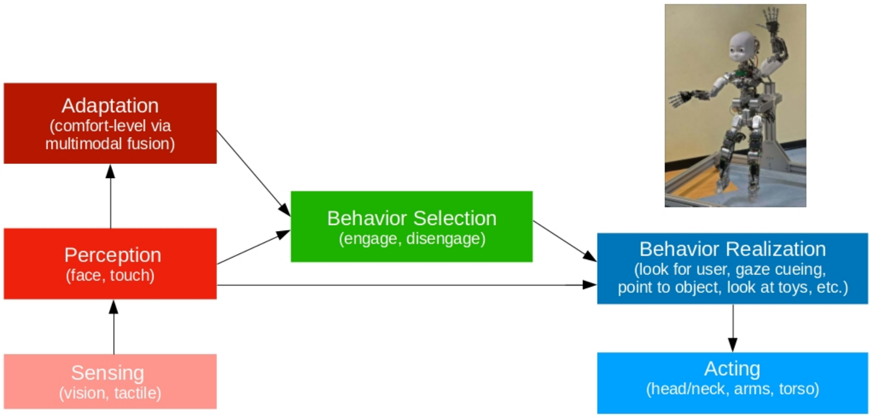


Figure 16.7 Single-route architecture (based on [Tanevska et al. 2019]) used to allow the iCub robot to adapt its behavior to the human over time (Inset: iCub robot [Metta et al. 2008]).

according to the information perceived from sensory data (e.g. position of the face). As can be seen, in this architecture, information flows from multimodal behavior processing to multimodal behavior generation modules and involves no high-level deliberation or planning.

Early efforts by Breazeal and colleagues [Breazeal et al. 2004] focused on providing social robots with key social competencies such as establishing joint attention during interaction with a human. For this, they developed an attention system that determines (1) what the human and the social robot, Leonardo, are looking at ("attentional focus") and (2) which objects they are referring to ("referent focus") during the interaction. To detect the referent focus of the human, information obtained through multiple perception modalities, especially speech and vision, are integrated (see Fig. 16.8). The visual information used includes pointing gestures made by the human and the eye gaze computed from the head pose. Information about the attentional and referent focus of the human and the robot are stored in the belief system, along with the attributes of the objects communicated by the human via speech. Updates to the beliefs (especially, the focus of the human and the robot) triggers several social behaviors, e.g. Leonardo would shift its gaze to the location that the human is currently looking at or point to the object being referred to. The architecture used by Breazeal et al. [Breazeal et al. 2004] (see Fig. 16.8) to enable such social behaviors is also a single-route architecture involving behavior/action selection based on beliefs held by the robot.

Dual-route architectures The dual-route architectures for social robots also involve a flow of information from behavior processing to behavior generation modules, but includes an *associative* and a *deliberative* route. These architecutures should support mechanisms to arbitrate and coordinate the behaviors generated via the two routes. The Cognitive and

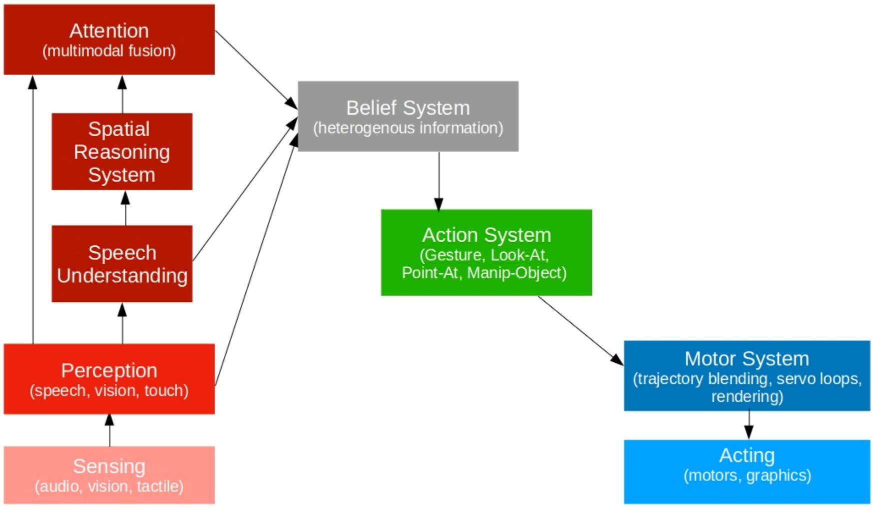


Figure 16.8 A single-route architecture based on [Breazeal et al. 2004], re-visualized according to our schematic. This architecture was developed for the Leonardo robot in order to achieve joint attention between a human and the robot.

Affective Interaction-Oriented Architecture (CAIO) proposed by Adam et al. [Adam et al. 2016] is an example of a dual-route architecture (see Fig. 16.9). It is focused on enabling social robots to reason about and express its affective state while also performing conversational acts simultaneously. The *dual-route processing* is facilitated by the sensorimotor and cognitive emotional appraisal modules (see Chapter 10 on "Emotion" [Broekens 2021] of volume 1 of this handbook [Lugrin et al. 2021] for a discussion of emotion models). The sensorimotor emotional appraisal module maps the conversational acts of the human interaction partner into a 5-D emotional representation, each of whose dimensions is mapped to specific face and body expressions by the Multimodal Emotional Action Renderer. This provides for a relatively short and fast *associative* route to express the initial emotional response of the robot to the human. The cognitive emotional appraisal module infers complex emotions (e.g. gratitude, reproach, etc.) for the robot based not only on the conversational act performed by the human, but also on the mental states of the robot. The complex emotions feed into a long and slow *deliberative* processing route, which involves the selection of the next intention for the robot, the composing of a plan of actions (conversational acts) to achieve the intention, and the execution of these actions through verbal and non-verbal modalities. The sensorimotor emotional appraisal module also evaluates the emotion associated with the conversational acts being executed by the robot, causing the expressed deliberative behavior to be emotionally flavored.

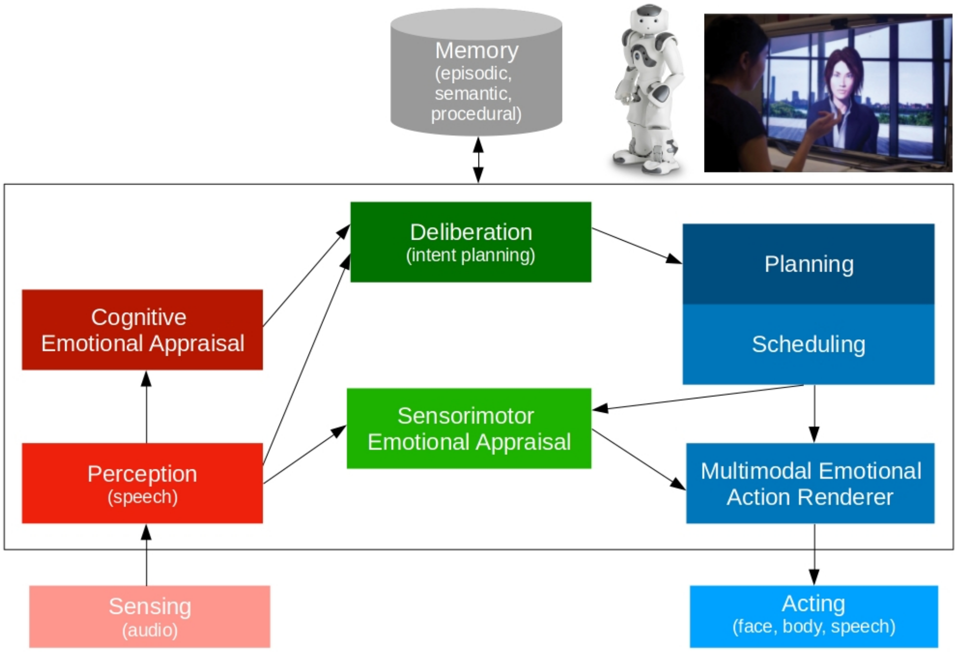

Memory
(episodic, semantic, procedural)
Deliberation
(intent planning)
Planning
Scheduling
Cognitive Emotional Appraisal
Sensorimotor Emotional Appraisal
Perception
(speech)
Multimodal Emotional Action Renderer
Sensing
(audio)
Acting
(face, body, speech)

Figure 16.9 CAIO Architecture (based on [Adam et al. 2016]), visualized according to our schematic, as another example of the double-route architecture that is employed in virtual agents and in social robots (Inset: (left) NAO robot © 2021 SoftBank Robotics; (right) MACH virtual conversation coach [Hoque et al. 2013]).

Multi-directional, incremental architectures Most social robot architectures developed so far focus mainly on multimodal behavior processing (red-colored boxes). Despite a lot of effort spent on the appropriate design of SR, the automatic generation of multimodal behaviors (blue boxes) remains an area that has received relatively little attention. Yet, naturalistic interactions require social robots to support the generation of interruptible, fluent and spontaneous multimodal behaviors (Sect. 16.1.1). That is, as with IVA, we need incremental, multi-directional architectures, which support incremental processing and exchange of information between components at all levels vertically, horizontally and diagonally (Sect. 16.2.1. Such architectures would (i) contribute desirable features like priming and anticipation, which can fasten and improve the reliability of behavior processing, and (ii) integrate multiple parallel routes for behavior generation (reactive, associative, deliberative) that operate at different temporal granularity [Kopp et al. 2014]. Recent work focuses on the development of a multi-directional SR architecture (see Fig. 16.10) that builds on an incremental communication framework [Schlangen et al. 2010], which was originally created for enabling naturalistic di-

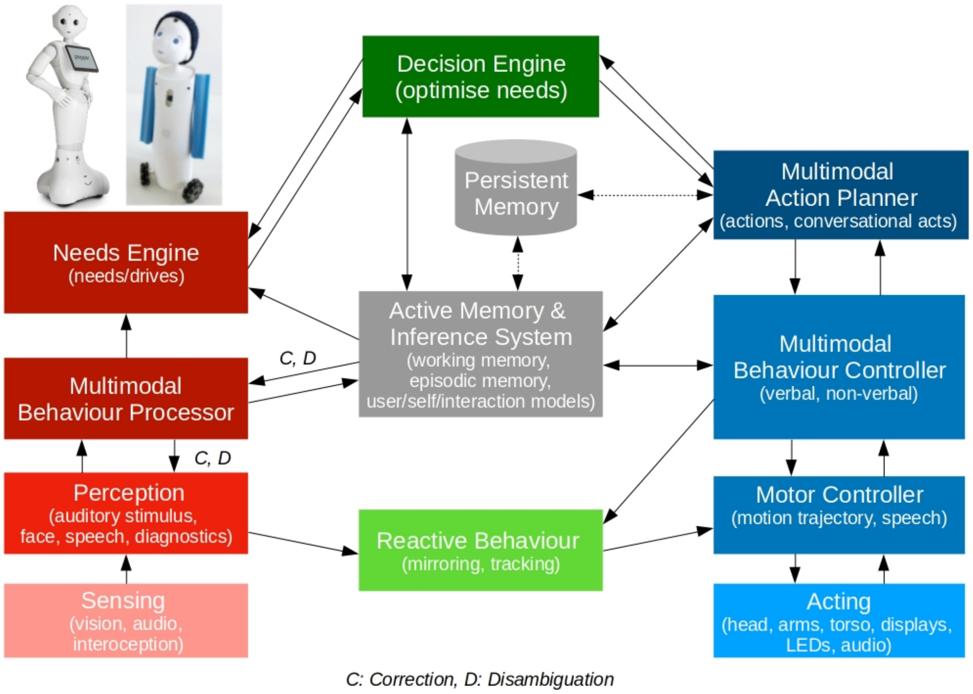

Decision Engine
(optimise needs)
Persistent Memory
Multimodal Action Planner
(actions, conversational acts)
Needs Engine
(needs/drives)
Active Memory & Inference System
(working memory, episodic memory, user/self/interaction models)
Multimodal Behaviour Controller
(verbal, non-verbal)
C, D
Multimodal Behaviour Processor
C, D
Perception
(auditory stimulus, face, speech, diagnostics)
Reactive Behaviour
(mirroring, tracking)
Motor Controller
(motion trajectory, speech)
Sensing
(vision, audio, interoception)
Acting
(head, arms, torso, displays, LEDs, audio)
C: Correction, D: Disambiguation

Figure 16.10 Example of a multi-directional, incremental architecture for enabling lively interactions between a human and a social robot (based on [Hassan and Kopp 2020, Stange et al. 2019]), visualized according to our schematic (Inset: (left) Pepper robot © 2021 SoftBank Robotics; (right) VIVA robot © 2021 navel robotics).

alog management in conversational virtual agents. The architecture was briefly introduced in [Stange et al. 2019] with a focus on the support for generating verbal explanations for behaviors, and in [Hassan and Kopp 2020] with a focus on the structure of its episodic memory.

The aforementioned architecture is being developed as part of a research project aimed at creating lively SIA for long-term interaction. For this, the architecture (Fig. 16.10) supports the dynamic modeling of intrinsic needs of the robot (Needs Engine) as well as the inference of the mental states of the user (User Model) powered by multimodal behavior processing. An elaborate incremental, multimodal behavior generation pipeline is included to fluently integrate behaviors at three conceptual levels: (i) fast, reactive behaviors (e.g. mirroring of facial expressions, tracking a human face, etc.); (ii) previously learned associative behaviors (e.g. idling when not engaged with the user, performing daily rituals such as greeting the user in the morning, etc.); and (iii) deliberately planned behaviors (e.g. getting acquainted with a new user, explaining own behavior to the user, etc.). A decision-making module (Decision

Engine) is used to select high-level intents aiming at optimizing the internal needs of the robots as well as those of the user. The flow of information is supported in multiple directions: bottom-up, top-down, left-right, and right-left. For example, feedback about the execution status of actions is used to adapt the selection of high-level behaviors in the future (bottom-up); the intrinsic needs of the robot are influenced by internal and external events perceived by the robot (bottom-up) as well as the success of its own plans (right-left); the discourse context as well as other information in the active memory are used to incrementally update/correct perception models and resolve ambiguities in perception and interpretation of sensor data (top-down); multimodal behavior controller purges/adapts certain reactive behaviors (top-down, right-left); active memory triggers idle behaviors according to the active interaction mode (left-right); dialog planner uses information gathered from past interactions to adapt future conversational acts (left-right). These illustrate the rich possibilities that the proposed multi-directional, incremental architecture would be able to provide to social robots in order to make them more lively, fluent and naturalistic in their interactions with the user. This architecture can be used not only for SRs but also for IVAs.

16.2.3 Similarities and Differences of Virtual Agent and Social Robot Architectures

Comparing the multimodal interaction architectures that have been developed and employed in virtual and robot agents, a number of commonalities but also differences can be noted. Overall, both kinds of systems have to implement the main columns for processing, mapping and generation, and to integrate them in a full architectural layout. In both fields, we find dual/multi-route architectures, which have been developed early on in the field of mobile robotics and have been adopted also for ECAs and social robots. Likewise, sub-architectures for online behavior processing or offline behavior generation have been developed and applied to both kinds of agents (e.g. [Ishi et al. 2018]). Also, multi-directional and incremental processing has been identified as an over-arching key feature and is addressed in architectures for both physical and virtual SIAs [Kopp et al. 2014, Stange et al. 2019].

However, a number of differences remain and hence offer opportunities for how one field can learn from the other. Obviously, one key difference is the physical embodiment of a robot and the constraints it implies for multimodal behavior processing (e.g. limited abilities to gather sensory information about a user's communicative behavior in a dynamic physical environment) as well as generation (e.g. limited abilities to produce expressive, subtle nonverbal behaviors under given bodily or kinematic limitations). Consequently, social robot architectures usually have concentrated on dealing with recognition problems (left-hand side of the schematic model) as well as fast lower-level routes in order to achieve situation-awareness and robust behavior. Social robots thus provide a great test-bed for embodied approaches to multimodal communication and socially reciprocal behavior coordination. In addition, although social robot bodies are carefully designed for expressiveness and engagement, their inherent physical limitations imply interesting challenges for researchers working on real-time generation of consistent, synchronized multimodal behavior [Ng-Thow-Hing et al. 2010]. Here, the transfer and application of behavior realization frameworks from virtual agents to social robots has led to some advanced generation sub-architectures that rest on closer feedback loops and more flexible timing and motion planning [Niewiadomski et al. 2013, Salem et al. 2012].

Full-fledged conversational social robots have been rarely reported in comparison to conversational virtual agents. Instead, the focus in SR research has so far been to explore processing and control mechanisms required for specific and by design pre-structured interactions between a human and a robot. Natural conversational interactions with, e.g., fluent turn-taking is still a challenge of such systems [Skantze 2020]. For example, Adam et al. [Adam et al. 2016] qualitatively demonstrated the integration of emotional appraisal with deliberately chosen conversational acts in order to produce multimodal (speech and gesture) behavior of the NAO robot [SoftBank-Robotics 2021a] based on an architecture that was initially developed for a virtual character. Even though Adam et al. [Adam et al. 2016] demonstrated how fast and slow cognitive processes could be integrated (see Fig 16.9), the interruptability of the deliberative or sensorimotor loops e.g. due to new incoming verbal input, was not explored. This requires a control strategy for managing turn-taking during conversations. Chao and Thomaz [Chao and Thomaz 2013] proposed the use of Timed Petri Nets (TPN) to regulate the conversational floor and thereby handle the dynamic turn-taking process in dyadic interactions between a human and a robot. Learning from and adapting to a user's preferences and skills is a key requirement for social robot companions. Park et al. [Park et al. 2019] used multimodal affective cues from verbal and non-verbal channels as 'reinforcement' or human feedback to adapt the storytelling policy of their social robot Tega, to increase the child's engagement with the robot and to improve the learning outcomes. These models could be directly applied to IVAs as well.

16.3 Current Challenges and Future Directions

In our discussion of the requirements of multimodal conversational interaction and the architectures used to build SIAs that shall be able to engage in it, we have already identified a number of trends and challenges. Given that the field is relatively young and still exploring new methods for behavior processing, mapping and generation, these challenges can be expected to persist for the next decades. In addition, a number of challenges and future directions can be identified that have or most likely will become crucial and the focus of this upcoming research.

Interaction memories and learning A core component of modern, learning-based architectures is a *memory* that aggregates information and makes it available for interaction with self and others. Memories can be distinguished according to what information they encode, how much, and how long they can retain it. Technical agent architectures differ in the types of memory systems that are included, their representations, the processes that operate on them, and how the memory influences other components and processes modeled within the architecture. For example, SOAR [Laird 2008, 2019] and MLECOG [Starzyk and Graham 2017] include short-term as well as long-term memory (episodic, semantic, procedural), while working memory plays an important role in the EPIC architecture [Kieras and Meyer 1997]. Likewise, cognitive architectures of SIAs often include memory as a key component. For example, CAIO stores information as long-term episodic, semantic, and procedural memories, but does not explicitly include a working memory. In [Dodd and Gutierrez 2005], short-term and long-term memories are used for the ISAC robot, and that includes sensory, episodic, semantic, and procedural memories. The architecture proposed in [Malfaz et al. 2011] use a long-term memory for supporting deliberative functions and a short-term memory for storing temporarily relevant information in the social robot Maggie [Salichs et al. 2006]. Kasap and Magnenat-Thalmann [Kasap and Magnenat-Thalmann 2010] propose long-term and short-term episodic memories to enable affective interaction between humans and social robots. The virtual agent SARA stores the adopted conversational strategies in “social history” and the preferences of and rapport with the user in the “user model” [Matsuyama et al. 2016].

Due to the increasing relevance of learning and adaptation in SIAs, the episodic memory is receiving growing attention, especially to store and provide access to past experiences and events. In agent architectures, episodic memories are usually created by filling pre-defined templates with specific information (cf. (Rabe and Wachsmuth, 2013), [Kasap and Magnenat-Thalmann 2010]) or by storing sequences of events that occurred while performing tasks (cf. [Dodd and Gutierrez 2005, Kasap and Magnenat-Thalmann 2010]). These models are quite restrictive, since the dynamism and complexity of interaction contexts make it difficult to predict the events that might occur during an interaction episode or the exact time at which they might occur. Nuxoll and Laird [Nuxoll and Laird 2007] proposed a design space to guide technical implementations of episodic memories. This could serve as a useful guide for current and future research on episodic memory models. In [Hassan and Kopp 2020], we presented a concept for an episodic memory model for storing interaction episodes, which addressed three aspects of this design space, namely when an interaction episode is encoded, what its content are and how it is structured. The proposed model represents episodic memory as hierarchies of labeled time-intervals when a user and an agent were engaged in an active interaction, initiated by either of the two parties. Relevant internal and external events are linked to the episodes based on their relationship with the episodes (i.e., causal, goal, or enabler events). A complete episodic memory model would however require that all aspects related to the encoding, storage and retrieval of episodic memories are addressed.

Cognitively inspired vs. application-centered architectures Related to the previous topic, a larger issue for future work will be to identify principles of cognitive architectures that facilitate social interaction, and to develop them into technical interaction architectures. Over the past decades, several architectures have been developed that identify, model, and weave together different cognitive processes in order to provide artificial agents (virtual agents or physical robots) with the computational framework for autonomous and intelligent behavior. Cognitive-psychological architectures include e.g. the widely known cognitive architectures ACT-R [Anderson et al. 2004], SOAR architecture [Laird 2008, 2019], and CLARION [Sun 2007] as well as the recently developed MLECOG [Starzyk and Graham 2017], which focuses on motivated learning capabilities. These architectures are generic and have a strong theoretical basis, but they do not focus on social reasoning or the generation of socially appropriate behavior that evolves over time. More application-centered architectures are usually developed to meet the requirements and demands of specific use-cases or applications, without giving much regard to cognitive or psychological principles. These solutions are frequently met in virtual agents as well as robotic systems and include, e.g., the architectures developed in [Kasap and Magnenat-Thalmann 2010] and [Dodd and Gutierrez 2005]. However, a common framework for an architectural layout that meets the requirements for fluent multimodal human-agent interaction and, in particular, integrates the many components that are needed on a principled basis is lacking. In particular, the increasing importance of integrating perception, action, memory, and learning is likely going to raise a need for cognitively plausible architectural concepts. For example, the need for robust and fluent interaction capabilities will require concepts for incremental yet concurrent and integrated processing at different layers of the architecture. Cognitive principles like good-enough reasoning or computational rationality may have a major play to role in these future systems.

Interaction-aware behavior generation Another important direction for the future, which is already starting to emerge, is the consideration and integration of the larger interaction context into local processing sub-architectures. While many works have investigated how background information such as personality, relational status or cultural background can be taken into account when processing or generating multimodal behavior (c.f. Chapters 7 on "Gesture Generation" [Saund and Marsella 2021], 8 on "Multimodal Behavior Modeling for Socially Interactive Agents" [Pelachaud et al. 2021], 13 on "Culture for Socially Interactive Agents" [Lugrin and Rehm 2021] of volume 1 of this handbook [Lugrin et al. 2021], and Chapter 18 on "Adaptive Artificial Person alities" [Janowski et al. 2022] of this volume of this handbook.), recent work has also started to explore how the current, dynamically changing interaction context can be integrated. While this has been often reported as crucial in human social interaction (e.g. for alignment, empathy or coordination), it has only recently and partially been tackled in technical attempts. For example, the current and previous body pose or facial expressions of the interlocutor have been integrated into the generation of respective behaviors of an agent [Ahuja et al. 2019, Jonell et al. 2020]. Those attempts are precursors of what we would term *interaction-aware multimodal behavior generation*. It will require new approaches to combine learning-based with model-based approaches that allow for learning at the level of interactional behavioral couplings and to embed this into the incremental processing of dynamic interactive behaviors.

Uncertainty-awareness in social interaction It is commonsense that systems that are to operate robustly in real-world environments, which are dynamic, stochastic, or only partially observable, require uncertainty-aware models. This is also true for SIAs, where uncertainty arises in the determination of the interaction context, in the inference of the meaning of social signals, in the attribution of mental states to the interaction partner, or in the prediction of possible effects of own multimodal behaviors. Uncertainty modeling is thus of special importance to the *processing* components of the multimodal interaction architecture shown in Fig. 16.1, since *sensing* is bound to be noisy, *recognition* models are not error-free, and *interpretation* is driven by hypotheses formed a under partial knowledge. Noise and errors also accumulate over successive processing stages in the architecture. This implies that the *mapping* components in Fig. 16.1 have to operate with uncertain information to decide, select or trigger an appropriate behavior of the agent. In the case of SRs, additional challenges for planning and control arise from the fact that *generation* components of the architecture have to deal with uncertain estimates of duration or outcomes of actions. Algorithms and approaches are available that can be used to handle uncertainties during interpretation (e.g. Bayesian networks for mental state modeling [Pöppel and Kopp 2018]) or decision-making (e.g. decision-theoretic planning methods). However, due to the complexity of these approaches as well as a lack of uncertainty estimates for multimodal *sensing* and *recognition*, existing SIA architectures often tend to ignore the handling of uncertainty during interaction. Recent efforts to quantify uncertainties associated with data-driven machine learning models and their predictions (cf. [Abdar et al. 2021]) – inspired by the seminal work of Yarin Gal (cf. [Gal and Ghahramani 2016, Kendall and Gal 2017]) – address this problem in perception tasks in real-world applications. Approaches for perception that combine prior knowledge with data-driven methods within probabilistic frameworks have also been reported recently (e.g. [Seuss et al. 2021]).

Future SIA architectures should focus on integrating uncertainty estimates into the interaction pipeline. That is to say, one should design, implement, evaluate, and optimize probabilistic versions of multimodal SIA architectures in order to promote their successful application in the real-world scenarios (outside laboratory settings). Furthermore, future human-robot interaction research should focus on investigating the influence of uncertainties on social interaction and on using probabilistic SIA architectures to autonomously generate such uncertainty-aware social behaviors. Finally, SIAs should be able to not only handle uncertainties technically, but also to create interactive behaviors that can adapt to as well as communicate various types and degrees of uncertainty.

Evaluation measures A final future direction that is going to become crucial is the development of metrics and criteria for the evaluation of multimodal behaviors. In the past and the present, the evaluation of multimodal behavior is mostly done by employing human raters in perception studies, who then rate the behavior for naturalness, coherence or style. There is a growing consensus in the field that an objective and more systematic evaluation methodology is missing and strongly needed (e.g. see [Wolfert et al. 2021] for a review). Further, such metrics are needed as optimization criteria for the development of machine learning-based models, which are increasingly trained in an adversarial fashion (i.e. using a discriminator or critic). Comparison against training data as usually done, however, is insufficient as the architecture's ability to generate multimodal behavior in new interaction contexts, in which it eventually needs to be communicatively effective and successful, cannot be assessed in this way. First approaches are seen for the evaluation of singular multimodal ensembles, e.g., by analyzing the internal temporal synchrony or semantic congruency between the modal behaviors. Yet, future work will need to investigate how a virtual agent's or social robot's long-term multimodal behavior in a given interaction context can be assessed. Finally, another aspect crucial in evaluating interaction architectures is how they can be inspected. An important future challenge is thus to increase the interpretability and transparency of an SIA architecture, such that developers as well as users can understand why a certain behavior has been shown. Indeed, the growing use of black-box machine learning models raises a need for explainable SIAs, whose interactive behavior can not only be received but also interrogated.

16.4 Summary

In this chapter, we presented the overarching requirements that should be fulfilled by virtual agent or social robot architectures in order to support fluent and human-expected multimodal interaction. We discussed several existing architectures for virtual and robot agents, especially with respect to the modalities they support, the different processing routes they use, and the aspects of social interaction they realise. We proposed a schematic for organizing the different modules and pathways in multimodal interaction architectures and categorized existing architectures into single-route, dual-route, and multi-directional architectures based on the processing pathways included. While the focus of virtual agent architectures has been mainly on expressive behavior generation, social robot architectures focused mainly on multimodal behavior processing and robustness. Either field is hence characterized by individual strengths and limitations. Although approaches are increasingly applied to both kinds of SIAs, the methods and models developed could inspire and inform each other even more to yield architectural principles and frameworks that enable advanced multimodal interaction capabilities to become standard in the field of SIAs.

Overall, the field has explored a variety of modalities, techniques, and integration architectures — and is still extending its repertoire of approaches for specific generation problems. Future work will need to consolidate our views on what we can generate, whether increasing data will help, how to join different approaches (and motivations), and what we should optimize models for (behavior quality metrics and felicity conditions). Future work should also focus on developing learning-based models that afford representations and interaction memories that can dynamically scale to heterogeneous data of different formats and temporal resolutions, and can enable an SIA to process, interpret, and learn from long-term interaction data in order to re-evaluate the social-appropriateness of behaviors based on past experiences.

Bibliography

M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76: 243–297. ISSN 1566-2535. https://www.sciencedirect.com/science/article/pii/S1566253521001081. DOI: https://doi.org/10.1016/j.inffus.2021.05.008.

C. Adam, W. Johal, D. Pellier, H. Fiorino, and S. Pesty. 2016. Social human-robot interaction: A new cognitive and affective interaction-oriented architecture. In A. Agah, J.-J. Cabibihan, A. M. Howard, M. A. Salichs, and H. He, eds., *Social Robotics*, pp. 253–263. Springer International Publishing, Cham. ISBN 978-3-319-47437-3.

C. Ahuja, S. Ma, L.-P. Morency, and Y. Sheikh. 2019. To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations. In *2019 International Conference on Multimodal Interaction*, pp. 74–84.

C. Ahuja, D. W. Lee, Y. I. Nakano, and L.-P. Morency. 2020. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *European Conference on Computer Vision*, pp. 248–265. Springer.

J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin. 2004. An integrated theory of the mind. *Psychological Review*, 111(4): 1036–1060. DOI: 10.1037/0033-295X.111.4.1036.

M. Atterer, T. Baumann, and D. Schlangen. 2009. No sooner said than done? testing incrementality of semantic interpretations of spontaneous speech. *Proceedings of Interspeech 2009*.

P. E. Baxter, J. de Greeff, and T. Belpaeme. 2013. Cognitive architecture for human–robot interaction: Towards behavioural alignment. *Biologically Inspired Cognitive Architectures*, 6: 30–39. ISSN 2212-683X. https://www.sciencedirect.com/science/article/pii/S2212683X1300056X. DOI: https://doi.org/10.1016/j.bica.2013.07.002. BICA 2013: Papers from the Fourth Annual Meeting of the BICA Society.

E. Bevacqua, K. Prepin, R. Niewiadomski, E. de Sevin, and C. Pelachaud. 2010. Greta: Towards an interactive conversational virtual companion. *Artificial Companions in Society: perspectives on the Present and Future*, pp. 1–17.

T. Bickmore, D. Schulman, and L. Yin. 2010. Maintaining engagement in long-term interventions with relational agents. *Applied Artificial Intelligence*, 24(6): 648–666.

A. Bono, A. Augello, G. Pilato, F. Vella, and S. Gaglio. 2020. An act-r based humanoid social robot to manage storytelling activities. *Robotics*, 9(2). ISSN 2218-6581. https://www.mdpi.com/2218-6581/9/2/25. DOI: 10.3390/robotics9020025.

T. Bosse, T. Hartmann, R. A. Blankendaal, N. Dokter, M. Otte, and L. Goedschalk. 2018. Virtually bad: a study on virtual agents that physically threaten human beings. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1258–1266.

C. Breazeal, A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, and D. Chilongo. 2004. Tutelage and collaboration for humanoid robots. *International Journal of Humanoid Robotics*, 1(2). DOI: 10.1142/S0219843604000150.

J. Broekens. 2021. *Emotion*, pp. 349–384. The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition. ACM Press. DOI: https://doi.org/10.1145/3477322.3477333.

H. Buschmeier, T. Baumann, B. Dosch, S. Kopp, and D. Schlangen. 2012. Combining incremental language generation and incremental speech synthesis for adaptive information presentation. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 295–303.

J. Cassell. 2001. Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine*, 22(4): 67–67.

J. Cassell, T. Bickmore, L. Campbell, H. Vilhjalmsson, and H. Yan. 2000. Human conversation as a system framework: Designing embodied conversational agents. In S. P. E. C. Justine Cassell, Joseph Sullivan, ed., *Embodied Conversational Agents*, chapter 2, pp. 29–63. MIT Press.

J. Cassell, H. H. Vilhjálmsson, and T. Bickmore. 2004. Beat: the behavior expression animation toolkit. In *Life-Like Characters*, pp. 163–185. Springer.

C. Chao and A. L. Thomaz. Feb. 2013. Controlling social dynamics with a parametrized model of floor regulation. *J. Hum.-Robot Interact.*, 2(1): 4–29. https://doi.org/10.5898/JHRI.2.1.Chao. DOI: 10.5898/JHRI.2.1.Chao.

C.-C. Chiu and S. Marsella. 2011. How to train your avatar: A data driven approach to gesture generation. In *International Workshop on Intelligent Virtual Agents*, pp. 127–140. Springer.

H. H. Clark and M. A. Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of memory and language*, 50(1): 62–81.

N. Crook, D. Field, C. Smith, S. Harding, S. Pulman, M. Cavazza, D. Charlton, R. Moore, and J. Boye. 2012. Generating context-sensitive eca responses to user barge-in interruptions. *Journal on Multimodal User Interfaces*, 6(1): 13–25.

W. Dodd and R. Gutierrez. Aug 2005. The role of episodic memory and emotion in a cognitive robot. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pp. 692–697. DOI: 10.1109/ROMAN.2005.1513860.

B. R. Duffy, M. Dragone, and G. M. O'Hare. 2005. Social robot architecture: A framework for explicit social interaction. In *Android Science: Towards Social Mechanisms, CogSci 2005 Workshop, Stresa, Italy*, pp. 3–4.

Y. Gal and Z. Ghahramani. 20–22 Jun 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In M. F. Balcan and K. Q. Weinberger, eds., *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059. PMLR, New York, New York, USA. http://proceedings.mlr.press/v48/gal16.html.

J. Gratch, J. Rickel, E. André, J. Cassell, E. Petajan, and N. Badler. 2002. Creating interactive virtual humans: Some assembly required. *IEEE Intelligent systems*, 17(4): 54–63.

T. Han, C. Kennington, and D. Schlangen. 2018. Placing objects in gesture space: Toward incremental interpretation of multimodal spatial descriptions. In *Proceedings of the AAAI Conference on Artificial*

Intelligence, volume 32.

Hanson-Robotics, 2007. Zeno. https://www.hansonrobotics.com/zeno/. Accessed: 2021-03-26.

D. Hasegawa, N. Kaneko, S. Shirakawa, H. Sakuta, and K. Sumi. 2018. Evaluation of speech-to-gesture generation using bi-directional lstm network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pp. 79–86.

T. Hassan and S. Kopp. 2020. Towards an interaction-centered and dynamically constructed episodic memory for social robots. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 233–235. Cambridge, UK. DOI: 10.1145/3371382.3378329.

A. Heloir and M. Kipp. 2010. Real-time animation of interactive agents: Specification and realization. *Applied Artificial Intelligence*, 24(6): 510–529.

A. Holroyd and C. Rich. 2012. Using the behavior markup language for human-robot interaction. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 147–148. IEEE.

M. E. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard. 2013. Mach: My automated conversation coach. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, p. 697–706. Association for Computing Machinery, New York, NY, USA. ISBN 9781450317702. https://doi.org/10.1145/2493432.2493502. DOI: 10.1145/2493432.2493502.

C. Huang and B. Mutlu. 2014. Learning-based modeling of multimodal behaviors for humanlike robots. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 57–64.

C. T. Ishi, D. Machiyashiki, R. Mikata, and H. Ishiguro. 2018. A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robotics and Automation Letters*, 3(4): 3757–3764.

K. Janowski, H. Ritschel, and E. André. 2022. *Adaptive Artificial Personalities*, pp. 155–193. The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application. ACM Press. DOI: https://doi.org/10.1145/3563659.3563666.

P. Jonell, T. Kucherenko, G. E. Henter, and J. Beskow. 2020. Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pp. 1–8.

Z. Kasap and N. Magnenat-Thalmann. Sep. 2010. Towards episodic memory-based long-term affective interaction with a human-like robot. In *19th International Symposium in Robot and Human Interactive Communication*, pp. 452–457. DOI: 10.1109/ROMAN.2010.5598644.

J. Kędzierski, R. Muszyński, C. Zoll, A. Oleksy, and M. Frontkiewicz. 2013. Emys–emotive head of a social robot. *International Journal of Social Robotics*, 5: 237–249. DOI: 10.1007/s12369-013-0183-1.

A. Kendall and Y. Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *CoRR*, abs/1703.04977. http://arxiv.org/abs/1703.04977.

D. E. Kieras and D. E. Meyer. 1997. An overview of the epic architecture for cognition and performance with application to human-computer interaction. *Human–Computer Interaction*, 12(4): 391–438. https://doi.org/10.1207/s15327051hci1204_4. DOI: 10.1207/s15327051hci1204_4.

S. Kopp. 2013. Gestures, postures, gaze, and movements in computer science: Embodied agents. Walter de Gruyter.

S. Kopp and N. Krämer. 2021. Revisiting human-agent communication: The importance of joint co-construction and understanding mental states. *Frontiers in Psychology*, 12: 597.

S. Kopp and I. Wachsmuth. 2004. Synthesizing multimodal utterances for conversational agents. *Computer animation and virtual worlds*, 15(1): 39–52.

S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, and H. Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *International workshop on intelligent virtual agents*, pp. 205–217. Springer.

S. Kopp, H. van Welbergen, R. Yaghoubzadeh, and H. Buschmeier. 2014. An architecture for fluid real-time conversational agents: integrating incremental output generation and input processing. *Journal on Multimodal User Interfaces*, 8(1): 97–108.

S. Kopp, M. Brandt, H. Buschmeier, K. Cyra, F. Freigang, N. Krämer, F. Kummert, C. Opfermann, K. Pitsch, L. Schillingmann, C. Straßmann, E. Wall, and R. Yaghoubzadeh. 2018. Conversational assistants for elderly users – the importance of socially cooperative dialogue. In E. André, T. Bickmore, S. Vrochidis, and L. Wanner, eds., *Proceedings of the AAMAS Workshop on Intelligent Conversation Agents in Home and Geriatric Care Applications*, CEUR Workshop Proceedings, pp. 10–17.

T. Kucherenko, P. Jonell, S. van Waveren, G. E. Henter, S. Alexandersson, I. Leite, and H. Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 242–250.

J. E. Laird. 2008. Extending the soar cognitive architecture. In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, pp. 224–235.

J. E. Laird. 8 2019. *The Soar Cognitive Architecture*. The MIT Press, Cambridge, MA, USA. ISBN 9780262538534.

J. E. Laird, K. R. Kinkade, S. Mohan, and J. Z. Xu. 2012. Cognitive robotics using the soar cognitive architecture. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

J. L. Lakin, V. E. Jefferis, C. M. Cheng, and T. L. Chartrand. 2003. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of nonverbal behavior*, 27(3): 145–162.

N. Leßmann, S. Kopp, and I. Wachsmuth. 2008. Situated interaction with a virtual human-perception, action, and cognition. In *Situated communication*, pp. 287–324. De Gruyter Mouton.

S. C. Levinson and F. Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6: 731.

M. Lhommet, Y. Xu, and S. Marsella. 2015. Cerebella: automatic generation of nonverbal behavior for virtual humans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

C. L. Lisetti and A. Marpaung. 2007. Affective cognitive modeling for autonomous agents based on scherer's emotion theory. In C. Freksa, M. Kohlhase, and K. Schill, eds., *KI 2006: Advances in Artificial Intelligence*, pp. 19–32. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-540-69912-5.

C. R. Ltd., 2020. MiRo - E. http://consequentialrobotics.com/miro-beta. Accessed: 2021-03-26.

B. Lugrin and M. Rehm. 2021. *Culture for Socially Interactive Agents*, pp. 173–211. The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition. ACM Press. DOI: https://doi.org/10.1145/3477322.3477336.

B. Lugrin, C. Pelachaud, and D. Traum. 2021. *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*. ACM Press. DOI: https://doi.org/10.1145/3477322.

M. Malfaz, Castro-Gonzalez, R. Barber, and M. A. Salichs. 2011. A biologically inspired architecture for an autonomous and social robot. *IEEE Transactions on Autonomous Mental Development*, 3(3): 232–246. DOI: 10.1109/TAMD.2011.2112766.

Y. Matsuyama, A. Bhardwaj, R. Zhao, O. Romeo, S. Akoju, and J. Cassell. 2016. Socially-aware animated intelligent personal assistant agent. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, pp. 224–227.

G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori. 2008. The icub humanoid robot: An open platform for research in embodied cognition. In *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, PerMIS '08, p. 50–56. Association for Computing Machinery, New York, NY, USA. ISBN 9781605582931. https://doi.org/10.1145/1774674.1774683. DOI: 10.1145/1774674.1774683.

C. Moulin-Frier, T. Fischer, M. Petit, G. Pointeau, J. Y. Puigbo, U. Pattacini, S. C. Low, D. Camilleri, P. Nguyen, M. Hoffmann, H. J. Chang, M. Zambelli, A. L. Mealier, A. Damianou, G. Metta, T. J. Prescott, Y. Demiris, P. F. Dominey, and P. F. M. J. Verschure. 2018. Dac-h3: A proactive robot cognitive architecture to acquire and express knowledge about the world and the self. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4): 1005–1022. DOI: 10.1109/TCDS.2017.2754143.

V. Ng-Thow-Hing, P. Luo, and S. Okita. 2010. Synchronized gesture and speech production for humanoid robots. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4617–4624. IEEE.

R. Niewiadomski, M. Mancini, and S. Piana. 2013. Human and virtual agent expressive gesture quality analysis and synthesis. *Coverbal Synchrony in Human-Machine Interaction*, pp. 269–292.

A. Nijholt, D. Reidsma, H. van Welbergen, R. op den Akker, and Z. Ruttkay. 2008. Mutually coordinated anticipatory multimodal interaction. In *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*, pp. 70–89. Springer.

A. M. Nuxoll and J. E. Laird. 2007. Extending cognitive architecture with episodic memory. In *AAAI*.

A. Papangelis, R. Zhao, and J. Cassell. 2014. Towards a computational architecture of dyadic rapport management for virtual agents. In *International Conference on Intelligent Virtual Agents*, pp. 320–324. Springer.

H. W. Park, I. Grover, S. Spaulding, L. Gomez, and C. Breazeal. Jul. 2019. A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 687–694. https://ojs.aaai.org/index.php/AAAI/article/view/3846. DOI: 10.1609/aaai.v33i01.3301687.

C. Pelachaud, C. Busso, and D. Heylen. 2021. *Multimodal Behavior Modeling for Socially Interactive Agents*, pp. 259–310. The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition. ACM Press. DOI: http://dx.doi.org/10.1145/3477322.3477331.

J. Pöppel and S. Kopp. 2018. Satisficing models of bayesian theory of mind for explaining behavior of differently uncertain agents. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*.

Robopec, 2021. Reeti: an expressive and communicating robot! http://www.reeti.fr/index.php/en/. Accessed: 2021-03-25.

M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin. 2012. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, 4(2): 201–217.

M. A. Salichs, R. Barber, A. M. Khamis, M. Malfaz, J. F. Gorostiza, R. Pacheco, R. Rivas, A. Corrales, E. Delgado, and D. Garcia. 2006. Maggie: A robotic platform for human-robot social interaction. In *2006 IEEE Conference on Robotics, Automation and Mechatronics*, pp. 1–7. DOI: 10.1109/RAMECH.2006.252754.

C. Saund and S. Marsella. 2021. *Gesture Generation*, pp. 213–258. The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition. ACM Press. DOI: https://doi.org/10.1145/3477322.3477330.

D. Schlangen and G. Skantze. 2011. A general, abstract model of incremental dialogue processing. *Dialogue & Discourse*, 2(1): 83–111.

D. Schlangen, T. Baumann, H. Buschmeier, S. Kopp, G. Skantze, and R. Yaghoubzadeh. 2010. Middleware for incremental processing in conversational agents. In *Proceedings of SIGdial 2010: the 11th Annual Meeting of the Special Interest Group in Discourse and Dialogue*.

D. Seuss, T. Hassan, A. Dieckmann, M. Unfried, K. R. R. Scherer, M. Mortillaro, and J.-U. Garbas. 2021. Automatic estimation of action unit intensities and inference of emotional appraisals. *IEEE Transactions on Affective Computing*, pp. 1–1. DOI: 10.1109/TAFFC.2021.3077590.

T. Shibata. 2012. Therapeutic seal robot as biofeedback medical device: Qualitative and quantitative evaluations of robot therapy in dementia care. *Proceedings of the IEEE*, 100(8): 2527–2538. DOI: 10.1109/JPROC.2012.2200559.

G. Skantze. 2020. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language*, p. 101178.

G. Skantze and A. Hjalmarsson. 2013. Towards incremental speech generation in conversational systems. *Computer Speech & Language*, 27(1): 243–262.

SoftBank-Robotics, 2021a. NAO the humanoid and programmable robot. https://www.softbankrobotics.com/emea/en/nao. Accessed: 2021-04-11.

SoftBank-Robotics, 2021b. Pepper the humanoid and programmable robot. https://www.softbankrobotics.com/emea/en/pepper. Accessed: 2021-04-11.

S. Stange, H. Buschmeier, T. Hassan, C. Ritter, and S. Kopp. 2019. Towards self-explaining social robots: Verbal explanation strategies for a needs-based architecture. In *Proceedings of the Workshop on Cognitive Architectures for HRI: Embodied Models of Situated Natural Language Interactions (MM-Cog)*. Montréal, Canada.

J. A. Starzyk and J. Graham. 2017. Mlecog: Motivated learning embodied cognitive architecture. *IEEE Systems Journal*, 11(3): 1272–1283. DOI: 10.1109/JSYST.2015.2442995.

R. Sun. 2007. The importance of cognitive architectures: an analysis based on clarion. *Journal of Experimental & Theoretical Artificial Intelligence*, 19(2): 159–193. https://doi.org/10.1080/09528130701191560. DOI: 10.1080/09528130701191560.

A. Tanevska, F. Rea, G. Sandini, L. Cañamero, and A. Sciutti. 2019. Eager to learn vs. quick to complain? how a socially adaptive robot architecture performs with different robot personalities. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pp. 365–371. DOI: 10.1109/SMC.2019.8913903.

C. Teufel and B. Nanay. 2017. How to (and how not to) think about top-down influences on visual perception. *Consciousness and cognition*, 47: 17–25.

M. Thiebaux, S. Marsella, A. N. Marshall, and M. Kallmann. 2008. Smartbody: Behavior realization for embodied conversational agents. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, pp. 151–158.

J. G. Trafton, L. M. Hiatt, A. M. Harrison, F. P. Tamborello, S. S. Khemlani, and A. C. Schultz. Feb. 2013. Act-r/e: An embodied cognitive architecture for human-robot interaction. *J. Hum.-Robot Interact.*, 2(1): 30–55. https://doi.org/10.5898/JHRI.2.1.Trafton. DOI: 10.5898/JHRI.2.1.Trafton.

D. Traum, D. DeVault, J. Lee, Z. Wang, and S. Marsella. 2012. Incremental dialogue understanding and feedback for multiparty, multimodal conversation. In *International Conference on Intelligent Virtual Agents*, pp. 275–288. Springer.

H. Van Welbergen, R. Yaghoubzadeh, and S. Kopp. 2014. Asaprealizer 2.0: The next steps in fluent behavior realization for ecas. In *International Conference on Intelligent Virtual Agents*, pp. 449–462. Springer.

H. Vilhjálmsson, N. Cantelmo, J. Cassell, N. E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A. N. Marshall, C. Pelachaud, et al. 2007. The behavior markup language: Recent developments and challenges. In *International Workshop on Intelligent Virtual Agents*, pp. 99–111. Springer.

P. Wolfert, N. Robinson, and T. Belpaeme. 2021. A review of evaluation practices of gesture generation in embodied conversational agents. *arXiv preprint arXiv:2101.03769*.

Y. Yoon, W. Ko, M. Jang, J. Lee, J. Kim, and G. Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 4303–4309. DOI: 10.1109/ICRA.2019.8793720.

Y. Yoon, B. Cha, J.-H. Lee, M. Jang, J. Lee, J. Kim, and G. Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6): 1–16.