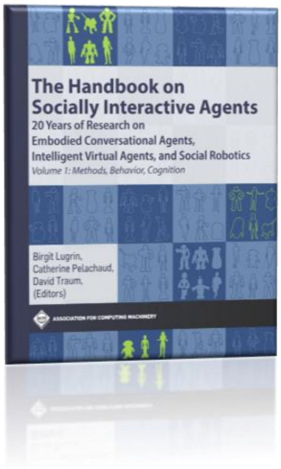
Multimodal Behavior Modeling for Socially Interactive Agents

Catherine Pelachaud, Carlos Busso and Dirk Heylen



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

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

ASSOCIATION FOR COMPUTING MACHINERY

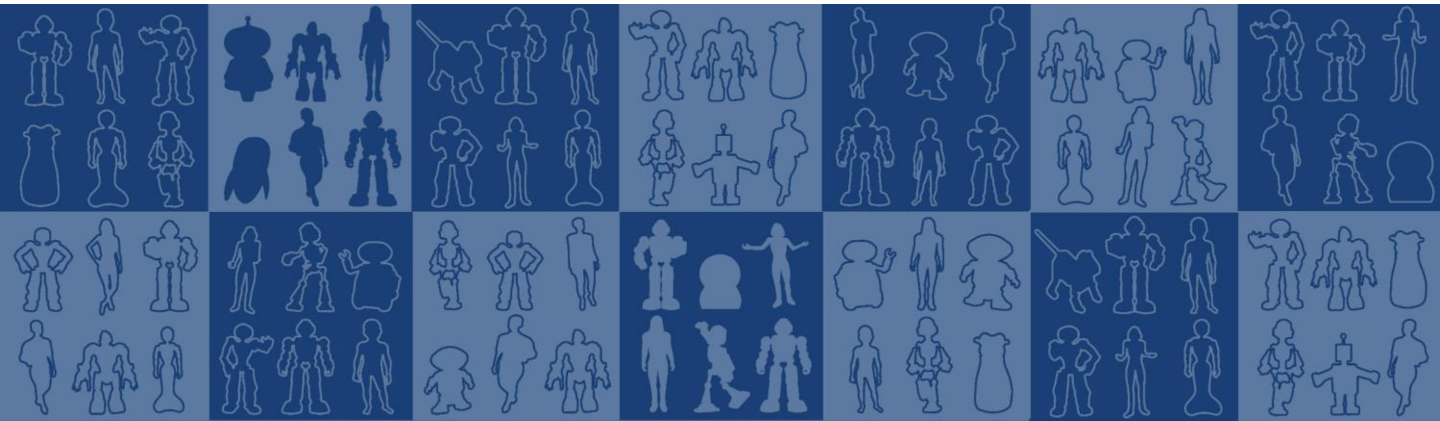
Citation information:

Pelachaud, C., Busso, C., and Heylen, D. (2021). Multimodal Behavior Modeling for Socially Interactive Agents. In B. Lugrin, C. Pelachaud, D. Traum (Eds.), *Handbook on Socially Interactive Agents – 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics*, Volume 1: Methods, Behavior, Cognition (pp. 259-310). ACM.

DOI of the final chapter: 10.1145/3477322.3477331

DOI of volume 1 of the handbook: 10.1145/3477322

Correspondence concerning this chapter should be addressed to Catherine Pelachaud (catherine.pelachaud@upmc.fr), Carlos Busso (busso@utdallas.edu) or Dirk Heylen (d.k.j.heylen@utwente.nl)



8 Multimodal Behavior Modeling for Socially Interactive Agents

Catherine Pelachaud, Carlos Busso, Dirk Heylen

8.1 Motivation

Imagine you start a conversation with the newest SIA, a humanoid robot or life size embodied virtual agent. It talks to you and you talk back, but the only parts moving are the lips. It does not smile when greeting you. There are no nods and shakes with its head and when it says "look over there" it does not point with its eyes or hands. You think to yourself: "What is wrong with it?" Is it working properly? Should I restart it? Or, who made this SIA? Did the makers not read the chapter on Multimodal Behavior in the Handbook on Socially Interactive Agents in which studies about the role of nonverbal communication in human-human communication are discussed and how such behaviors have been implemented in several generations of SIAs? Given that these SIAs are embodied conversational agents, it is expected that they mimic as best as they can the nonverbal behaviors that humans use in their face-to-face conversations. Take the case of gaze.

The eyes are the organs that - as part of our visual system - make us see the world and the people around us. We look at what attracts us or at objects that we manipulate and where vision helps us to successfully perform the action. Sometimes we might stare without looking at anything in particular; our mind lost in thought. Most of the time, however, we look because we need to pay attention to what is happening and because we need to act. We look at the world and the people around us for a reason. As do the people around us. As we can see the other looking at us, we might start wondering why we are being looked at and perhaps also what the other person thinks is the intention behind the reason why we are looking at her/him. We look at each other in different ways. We have learned and we have been taught what intents lie behind a glance or a stare and what conventions govern interactions in particular situations. We know how somebody tries to direct our attention to something by pointing to it with her/his eyes. We look at the specific person that we are addressing when we are together in a group and know that we are being talked to when somebody talks and looks at us. From the other side, we pay attention to the person who is talking to us by looking at the person.

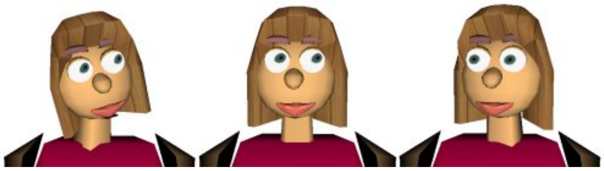


Figure 8.1 Different ‘look-away’ behaviors [Heylen et al. 2005]

All these behaviors help us to regulate the flow of interaction. We have also learned to see how somebody is angry at us by a fierce stare and when somebody wants to seduce us with a flirtatious look - accompanied by a slight head tilt and a squeezing of the eyes.

The same kind of interpretations apply when we look into the eyes of an embodied SIA as Fukayama et al. [2002] demonstrated by showing interface agents that merely consisted of eyes that displayed different gaze patterns. By changing parameters such as the amount of gaze and the mean duration of gaze, they showed how the eyes conveyed different impressions related to liking, warmth and potency.

Based on the suggestions to link gaze patterns with turn-taking and discourse structure by Torres et al. [1997], Heylen et al. [2005] experimented with an agent displaying different gaze patterns that were designed to provide smooth interaction patterns by providing the right turn-taking cues (see Figure 8.1). Besides showing that using the appropriate patterns made a difference in task performance they also showed an effect on the trustworthiness of the agent and the appreciation of warmth. This goes to show that simple nonverbal behaviors have multiple effects and are important to get right not only in order to be efficient and effective, but also to have the agents give the desired impression in terms of emotion and interpersonal relationship as this will have an effect on how someone will want to keep being engaged with the agent.

Gaze is just one of the modes of nonverbal communication that one needs to pay attention to in building SIAs, as we will see in this chapter. Facial expressions, head movements, gestures and posture, all play an important part in the impressions the agents convey and the way a person engages with them (see Figure 8.2).

In this chapter, we present studies on the generation of multimodal nonverbal behaviors for SIAs. We first present some works on nonverbal communication in human human interaction and next move to computational models. In the History section, we present a variety of topics that received attention at a particular time. We go in some depth on machine learning approaches to the generation of nonverbal behaviors, after which we conclude the chapter by discussing some current and future challenges.

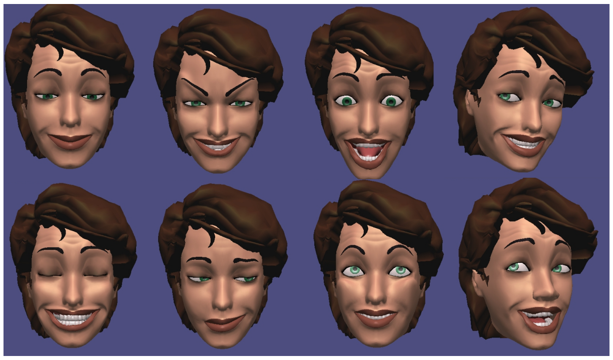


Figure 8.2 **Different gaze and facial behaviors (adapted from [Heylen 2010])**

8.2 Nonverbal behavior representation

In this section, we start by presenting different taxonomies of nonverbal behaviors as well as coding schemes that have been proposed in the literature of human and social sciences.

8.2.1 Classification of nonverbal behavior

As introduced above, nonverbal behaviors convey a variety of meanings. To study them, scholars have proposed to cluster them depending on their characteristics. Several attempts have been made to classify them depending on their communicative, emotional or pragmatic functions. Classifications may be geared toward one modality (cf gesture taxonomy by McNeill [1992], facial expression taxonomy by Chovil [1991]) or encompass multimodality (e.g. Bavelas et al. [2014], Ekman [2004] or Poggi [2007]).

Ekman [2004] distinguishes between facial and body movements used as emotional or conversational signals. He refers to the latter ones as illustrators since the signals *illustrate* speech. Signals such as hand movements or brows movements often accompany a change of loudness in the voice. Deixis refers to indicating a point in space; it is also part of the illustrator cluster. Deixis can be conveyed by an extended hand gesture, a head and eye direction, and even by a chin upward movement. Pictographs that embed signals imitating an object (referred to as iconic by other scholars [McNeill 1992]). Another other cluster gathers *emblem*, including *emotional emblem* that are produced to replace common verbal expressions, and correspond to movements whose meanings are very well-known and culturally dependent (e.g. nodding instead of saying “yes”). The other two clusters embrace manipulators (such as touching one’s face) and regulators (to maintain and regulate speaking turns) and emotional expressions.

McNeill [1992] focuses mainly on communicative gestures. He classified them by the type of information they convey. There are the *iconics* where hand gesture describes a concrete object, the *metaphorics* for which motion of the hand represents an abstract idea, the *deictic* where finger pointing designs a point in space, *beat* that corresponds to rhythmic movement of the hand, and *emblem* that includes gestures whose meaning are culturally coded.

Another example of taxonomy has been proposed by Poggi [2007]. She characterizes behaviors based on the semantic functions they convey [Poggi 2007]. She considered three main categories, each of them divided into sub-categories. Speakers may communicate information about the world (e.g. deictic and iconic gestures), one's own identity (age, culture, personality, gender) and the mind (belief (certainty etc.), goal (performative, intonational structure...), and emotion). To capture the polysemy of nonverbal behaviors, she defines each element of these categories by a pair where the first element is the meaning and the second one lists the multimodal signals that convey this meaning. As such, the same signal may be present in different pairs and linked to different meanings.

As we can see in these few examples, taxonomies show similarities and differences. They differ in the type of information that are used to distinguish signals or on the granularity of the taxonomy. Moreover, before choosing a taxonomy to cluster behaviors, we would like to raise some concerns. Indeed, one has to be aware with a simplification that may arise when using taxonomies. The same behavior may have different meanings depending on its context of occurrence – e.g. a smile can be a sign of happiness or politeness, or serve as a backchannel; a hand moving upward can be an iconic gesture and can also be considered as a deictic gesture. Identically, a communicative function may be expressed by various signals – to communicate refusal, one can shake the head or shake the index finger.

Context is a crucial factor when interpreting a social signal. However it is not encoded in the taxonomies. Some scholars, in particular Bavelas and Chovil [2000], warned also about using taxonomy. Taxonomy defines classes that are mutually exclusive, which is incompatible with polysemous characteristics of nonverbal behaviors. Poggi [2007] takes into account the polysemy of nonverbal behaviors, but her taxonomy does not include any contextual information. Rather than defining a taxonomy, Bavelas and Chovil [2000] proposed to use the term "visible acts of meaning" and to include behaviors that are visible, communicative and linked to the on-going speech. They highlight the need to look at nonverbal behaviors by their communicative functions in the context of the speech.

8.2.2 Coding system

To be able to discuss about a behavior, one needs to be sure the term being used is commonly understood by everyone and everyone attaches the same meaning to it. Let us consider smile as an example. Smile corresponds to a well defined facial expression often associated to happiness. However smiles can take very diverse forms associated with a large number of meanings. Just think of the smile of the Joker or of sportswomen/men arriving at the second place, or even of politicians; they smile but their smile differ greatly in their appearance and communicative functions. So the label “smile” is confusing. It cannot be associated to a common meaning.

To analyze and describe nonverbal behaviors, it is important to have a common representation language. To this aim, several coding schemas have been proposed; some focus on one modality, others cover multiple ones. We list here the most popular ones. The different schemes differ by the modality they focus on and by their granularity.

The Facial Action Coding System (FACS) was developed by Ekman, Friesen and Hager to describe facial expressions [Ekman et al. 2002]. The authors defined an Action Unit (AU) as the minimal muscular contraction that is visible. A facial expression corresponds to the combination of AUs. 44 AUs are defined divided along the facial regions. An expression is also defined by its temporal course, where onset is the time of appearance of an expression, offset the time of disappearance and apex is the time where the expression is at its maximum. There exist variants of FACS such as EM-FACS for facial expressions of emotion [Friesen et al. 1983] and babyFACS for expressions of babies [Oster 2006].

Several annotation schema have been designed to code body expression of emotions. The Body Action Coding System (BACS) proposed by Huis In‘t Veld et al. [2014] follows a similar concept as FACS. Contrary to Kendon [1980]’s approach that describes the shape and the movement that constitute a gesture, BACS describes the muscular activation, measured through electromyogram (EMG), involved in a movement. The approaches capture very complementary information. The aim of BACS is to describe body expressions of emotions. Other schemas have been proposed to describe body movement on a multi-level labeling approach [Dael et al. 2012, Fourati and Pelachaud 2014]. Dael et al. [2012] have developed the Body Action and Posture coding system BAP. BAP was designed to describe body movement of emotions that is visible. It describes body movement along three main levels: 1) anatomical level working at the body articulation level; 2) form level encoding direction and orientation of body movement; and 3) functional level that corresponds to communicative and self-regulatory functions. BAP also offers mechanisms to describe temporal relationship between body parts. Fourati and Pelachaud [2014] proposed also a multi-level annotation scheme to describe body actions. Their scheme encompass descriptions at anatomical level, directional level, and posture/movement level.

Regarding communicative gestures, to our knowledge, there is not an annotation scheme that is commonly used among scholars (see Chapter 7 on communicative gestures).

However, often a communicative gesture will be defined by the position of the wrist in 3D space, palm orientation and fingers shape. Kendon [1980] defined the temporal course of hand gestures along different phases: preparation, pre-stroke, stroke, post-stroke, hold, retraction; where stroke is the phase that carries the gesture meaning. Between consecutive gestures that are close enough temporally, hand movement may co-articulate between the last phase of the precedent gesture into the first phase of the successive one. We can note that the classification and the decomposition in phases can be applied to other modalities such as facial expressions and head movements. For example, the Behavior Markup Language BML [Kopp et al. 2006] follows these ideas (see Chapter 16 on multimodal architectures).

Lately the modality touch is receiving more and more attention from the SIA community (see Section 8.4.6). To our knowledge there is no common annotation scheme that has been drawn for touch and used by various researchers. Instead, researchers made up their own annotation schemes that share several features. To define the conformational parameters of social touch gestures, Teyssier et al. [2020] and Atkinson et al. [2013] relied on human-human interaction studies [Hertenstein et al. 2009]. These parameters encompass the action of the touch gestures [Atkinson et al. 2013, Hertenstein et al. 2009] – for example hit, pat, stroke, caress, shake... to name a few. Each action can be further represented by the characteristics of the movement itself: velocity, amplitude, force of touch movement. Temperature of the body part giving the touch and of receiving it can also be encoded. It can also be applied to self-touch (e.g. often called adaptator by scholars [Ekman 2004]).

While the schemas we have just presented focus on one modality only, the annotation scheme MUMIN [Allwood et al. 2007] was developed to annotate multimodal communication. MUMIN has a multi-layer structure that encode low and high level information. The latter one corresponds to the list of signals of each modality. For example the item facial display (face modality) encompasses signals of five facial regions: *Eyes, Gaze, Eyebrows, Mouth, Head*. Each region can have different values. For e.g., *Eyebrows* can get 3 possible values: *Frowning, Raising, Other*; and *Gaze* 5 different ones: *Toward Interlocutor, Up, Down, Sideways, Other*. Communicative functions are annotated at the highest layer. Three main classes of communicative functions are considered: *Feedback (Give / Elicit), Turn-Managing* and *Sequencing*. Each class can take many forms that are instantiated into several values. To illustrate the different levels of granularity of this annotation scheme, let us consider the function *Feedback-Elicit*. It can be instantiated into three sub-classes: *Basic, Acceptance, Additional Emotion / Attitude*; each of these sub-classes can receive different values. Such a multimodal and multi-layers coding system allows learning about the role of gesture, facial expression, head... in conveying communicative functions. Different analysis methods, e.g. statistical or sequencing, can be applied to build models mapping communicative functions and multimodal behaviors.

Further information on coding schemes can be found in [Jokinen and Pelachaud 2013].

8.3 Models and Approaches

We present here the broad categories of models that have been proposed so far to model IVA's nonverbal behaviors.

Over the last decades, the research community has investigated different computational approaches to add nonverbal behaviors to SIAs. The early formulations were based on carefully designing *rules* to communicate particular discourse meaning. Examples of these approaches include the studies of Beskow and McGlashan [1997], Cassell et al. [1994, 1999, 2001], Kopp and Wachsmuth [2002, 2004], Marsella et al. [2013] and Poggi and Pelachaud [2000]. The rules were often obtained by learning relationships between communication channels. For example, these methods leverage what we know about the discourse functions of hand gestures [McNeill 1992], head motion [Heylen 2005, Sadoughi and Busso 2017b] and facial displays [Chovil 1991]. The text is often analyzed, selecting synchronization points to include nonverbal gestures to emphasize or clarify the message. For example, Kopp and Wachsmuth [2004] proposed to identify prominent words or phrases, which were used to anchor specific nonverbal gestures.

Performance-based nonverbal generation is another popular approach that aims to use concatenation of human recordings that are re-purposed to create the behavior in the animation [Kipp et al. 2007, Neff et al. 2008, Rizzo et al. 2004, Stone et al. 2004, Williams 1990]. The advantage of these approaches is that the original synchronization between speech and gestures is preserved, resulting in natural renditions. Examples of this technique include the work of Arikan and Forsyth [2002] and Lee et al. [2002], where motion capture recordings were concatenated to create a novel sequence. Performance-based generation has been used for lip synchronization, where frames are carefully concatenated to match the phone sequence [Bregler et al. 1997]. The key challenge with this approach is to generalize the system to new sentences that are not in the database. It is important to create flexible approaches to combine and smooth the transition between pasted sequences [Lee et al. 2002].

Recent studies have also proposed to incorporate nonverbal behaviors by using *machine-learning* approaches, obtaining the relationship across gestures directly from the data. For example, these approaches leverage the rich relationship between speech and gestures [Brand 1999, Busso and Narayanan 2007, Graf et al. 2002, Kettebekov et al. 2005, Valbonesi et al. 2002]. Studies to generate nonverbal behaviors have used graphical models, including hidden Markov models (HMMs) [Busso et al. 2007, Le et al. 2012], dynamic Bayesian model (DBN) [Mariooryad and Busso 2012, Sadoughi et al. 2017], hidden conditional random fields (HCRF) [Levine et al. 2010], fully parameterized Hidden Markov model (FPHMM) [Ding et al. 2013a] and hidden semi Markov models [Bozkurt et al. 2008]. Other machine-learning approaches to generate nonverbal gestures have relied on solutions based on Deep Neural Network (DNN). Examples of implementations of nonverbal behavior prediction models using DNNs include architectures with fully connected layers [Parker et al. 2017, Sadoughi and Busso 2018b, Taylor et al. 2016, 2017], long-short term memory (LSTM) [Fan et al. 2016, Li et al. 2016, Sadoughi and Busso 2017a], generative adversarial network (GAN) [Huang and Khan 2017, Sadoughi and Busso 2018a, 2020], and convolutional neural networks (CNN) [Karras et al. 2017].

Some studies have proposed *hybrid approaches* to combine rule-based systems with machine learning models. One approach is to constrain the data-driven model to a specific gesture (e.g., head nod), or discourse function (e.g., asking questions) [Sadoughi and Busso 2015, 2019, Sadoughi et al. 2014, 2017]. These hybrid methods have the advantage of providing novel realizations of the target gesture or behavior that can be intrinsically synchronized with other modalities (e.g., speech), preserving the meaning intended for the message. These models can operate between the *behavior planning* and *behavior realization* modules of the SAIBA architecture (for further details, see Chapter 16 presenting different agent architectures).

8.4 History / Overview

We now turn our attention to the presentation of existing works on modeling nonverbal behaviors for SIAs. Though, the focus of this section is primarily on IVAs' behavior models. We start from the earliest works and follow a somewhat chronological order. However, studies are also presented by the functions they target, such as the expression of emotion or the modeling of rapport. A pure historical presentation is not possible as the same topics have been addressed over the years but with very different methods.

8.4.1 Early works

In the 1970's the first facial models appeared [Parke 1972]. Research focused not only on modeling lip movement and co-articulation effects during speech [Beskow 1997, Cohen and Massaro 1993, Parke 1975], but also the facial expressions of emotion [Pelachaud et al. 1996]. The first system where two virtual agents entered into a dialogue together and showed nonverbal behaviors was *GestureJack* [Cassell et al. 1994] (see Figure 8.3). Gesture, facial expression, head movement and gaze were aligned automatically with speech. Different rules were designed to drive the agent's behavior. Rules came from the human and social sciences literature. They were also extracted from corpus analysis of human interactions. They specified the type of a behavior (e.g., an iconic gesture, a backchannel), its form (such as a writing gesture, a head nod), and its timing (when it starts and how long it lasts). The *GestureJack* system included also a dialog model as well as a voice synthesizer. It laid the foundation of Embodied Conversational Agents (ECA) as all the nonverbal behaviors for the agent acting as a speaker or as a listener were automatically generated.

Later on, Cassell and her colleagues developed the human-agent interaction system, called REA [Cassell et al. 1999] (see Figure 8.4). REA stands for Real-Estate Agent. Users could converse with REA using a spoken dialog. User's gaze direction and speech were extracted in real-time. To establish some form of rapport, the REA agent chatted with the user and exchanged small talk. It used multimodal behavior to communicate. For example, it could display iconic gestures to describe houses as well as interactional gestures to handle turn-taking. Turn-taking mechanisms were further developed by Thórisson [1997] in the Ymir architecture. The perception module included several modalities. Gaze behavior and deictic gestures of the user were detected and used to understand which object(s) the user was interested in discussing. Other agent models integrating verbal and nonverbal behaviors have been proposed by the late nineties [André et al. 1998, Pelachaud and Poggi 1998, Rickel

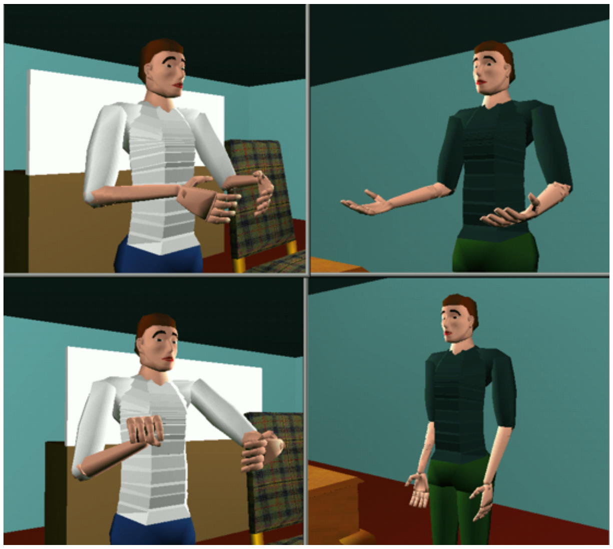


Figure 8.3 **GestureJack: first complete IVA system; two fully autonomous agents interacting together [Cassell et al. 1994].**

and Johnson 1999]. The agents were integrated into a system architecture that consisted of multi-layer components where one component made plans and took decisions and another one instantiated the latter onto nonverbal behaviors.

Communication involves multimodality where nonverbal behaviors are displayed with complex and sophisticated relations. Several approaches were developed at the turn of the century to capture these relations. Cassell and Stone [1999] designed a multimodal manager whose role was to supervise the distribution of behaviors across the verbal and nonverbal channels (verbal, face, gesture, head, and gaze). The Behavior Expression Animation Toolkit (BEAT) [Cassell et al. 2001] is a tool that can generate and synchronize facial expression and gestures from text (see Figure 8.5). Rules were designed from human behavior studies to indicate where and which gestures could plausibly fit. The text to be spoken by the agent was further decomposed into theme and rheme, that correspond, respectively, to known versus new information carried by the utterance [Halliday 1967]. It allows us to place gestures on the rheme, that is when new information of the utterance is said by the agent.

Poggi and Pelachaud [2000] considered some contextual information to instantiate communicative intentions into behaviors. They focused on the notion of performative, the illocutionary force of an utterance [Austin 1962], where they consider as context the relationship between the speaker and the listener. They applied a meaning-to-face approach to compute the appropriate facial expression driven by semantic analysis.

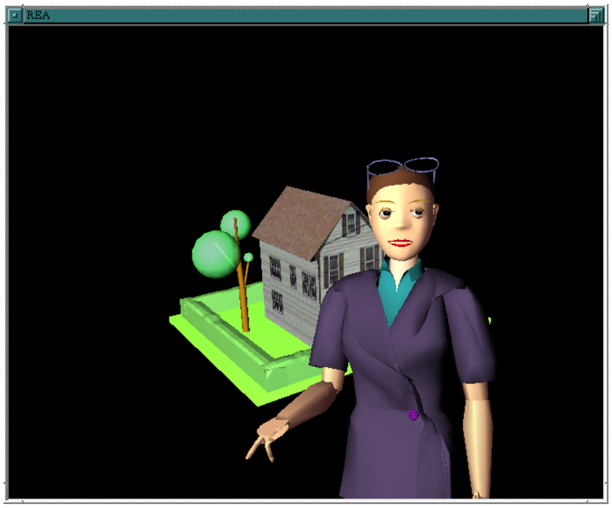
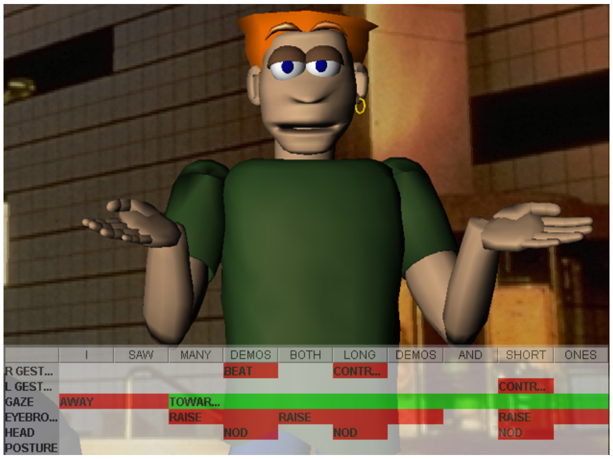


Figure 8.4 REA performs an iconic gesture.



	I	SAW	MANY	DEMOS	BOTH	LONG	DEMOS	AND	SHORT	ONES
R GEST...				BEAT		CONTR...				
L GEST...									CONTR...	
GAZE	AWAY		TOWAR...							
EYEBRO...			RAISE		RAISE				RAISE	
HEAD				NOD		NOD			NOD	
POSTURE										

Figure 8.5 Nonverbal behaviors computed by the BEAT system.

Deictic behaviors received also particular attention for the creation of pedagogical agents [André et al. 1996, Lester et al. 2000]. These agents would indicate objects in 2D or 3D space that were the focus of the discussion. Indicating a point in space can be done through a pointing gesture, walking to this point, gazing at it, turning one's head toward it or even doing a chin-up movement in its direction. Choosing the behaviors to display requires knowing where the object is in space in relation to the agent's position as well as if this object was already mentioned and thus known or not to the learner.

Lee and Marsella [2006] applied such a methodology when developing the Nonverbal Behavior Generation (NVBG) System. They annotated corpora on several levels: dialog acts (e.g. affirmation, obligation, listing), text and behaviors (head movement, eyebrow movement, gaze and other). These analyses allowed the authors to draw rules that mapped a dialog act onto instances of text and nonverbal behaviors [Lee and Marsella 2006]. The dialog acts are encoded with Function Markup Language (FML) [Heylen et al. 2008] and the output of the NVBG system in the Behavior Markup Language (BML) [Vilhjálmsson et al. 2007] of the Situation, Agent, Intention, Behavior, Animation (SAIBA) architecture [Kopp et al. 2006].

8.4.1.1 Expressions of emotions

Emotions are crucial in everyday life. Chapter 10 on emotion presents different theories in psychology that have led to computational models. The latter ones trigger an emotion or a blend of emotion the agent ought to convey through its choice of word, prosody, facial expression, gaze and body movement.

Static representations At first, researchers [Pelachaud et al. 1996, Thalmann et al. 1998] relied on the literature, and in particular on the so-called "six basic" expressions of emotion that Ekman claimed to be universally recognized and produced [Ekman and Friesen 1975]. The expressions are described in terms of AU from the coding system FACS [Ekman et al. 2002]. The standard MPEG-4 [Ostermann 2002] also supports the encoding of these expressions. Further computational models were proposed to create a great variety of facial expressions of emotion. They followed the dimensional approach promoted by the core affect theory (see Chapter 10 to have an overview of these approaches). Ruttkay et al. [2003] created the EmotionDisc, a disc acting as an interface where expressions of emotion were spread along with the neutral expression at its center. A new expression is obtained as the bi-linear interpolation of the facial expressions of two emotions that can have different intensities. This approach was extended to a 3D space. A new expression was computed as the linear interpolation of two neighboring expressions that are already defined [Albrecht et al. 2005, Tsapatsoulis et al. 2002]. Courgeon et al. [2008] also applied the interpolation between expressions of emotions defined in the three dimensional space of pleasure (P), arousal (A) and dominance (D) [Mehrabian 1996]. The authors made use of a joystick whose displacement between points (i.e. expressions) in the PAD space was used to compute the intermediate expressions between the points. These models view the facial expression as a whole. Other models proposed a compositional approach where a facial expression is defined as a set of regions and a new expression is computed as the combination of expressions in its different regions. This approach was used to create blends of expressions such as the superposition of two expressions or the masking of an expression by another one. Fuzzy rules were designed to determine the blending of expressions over the facial regions [Bui 2004, Niewiadomski and Pelachaud 2007]. Arya and DiPaola [2007] used fuzzy rules on perceptually validated believable expressions of emotion. Rehm and André [2005] modeled the expression of non-felt emotion as asymmetric and without the reliable features as defined by Ekman and Friesen [1975].

Temporal representations The previous models described expressions of emotion by their apex, where expressions are defined at the maximal muscular contractions. Other models viewed the expressions of emotion as temporal sequences of facial actions. While some of these models [Courgeon et al. 2014, Malatesta et al. 2009, Paleari and Lisetti 2006] link these sequences with appraisal checks [Scherer 2001], some models do not explicitly do so [Jack et al. 2012, Niewiadomski et al. 2011, Pan et al. 2007]. The former models have implemented how facial expressions of emotion are built in based on the appraisal model proposed by Scherer and Ellgring [2007]. Facial expressions arise from the temporal evaluation process of an event along a set of Stimulus Evaluation Checks SEC (see Chapter 10 on emotion). Scherer [2001] has defined a mapping between evaluation of these SEC and facial signals. The facial expression of an emotion is, thus, a sequence of dynamic facial signals. These computational models differ on how the facial signals are merged on the face; they can either be adding up on the face or the facial signals have a given temporal course. These computational models highlighted some missing information of the theoretical model (such as the intensity of facial signals or their duration). The studies by Niewiadomski et al. [2011] and Pan et al. [2007] relied on the analysis and annotation of videos of humans expressing emotions. Pan et al. [2007] built a motion graph where the arcs of the graph correspond to motion clips and the nodes to possible transition clips. A new sequence of facial signals can be created by choosing a different motion graph. After annotating facial action using FACS, head, gesture and body motion, Niewiadomski et al. [2011] designed a set of temporal and spatial constraints. These constraints were defined to capture the relationship between the multimodal signals. The expression of an emotion is a sequence of signals that respect these constraints (see Figure 8.6). An advantage of both these methods is the ability to create different sequences of behaviors while maintaining their meaning (the conveying of a given emotion). It allows the agent to display a greater variety of behaviors. Lately, Jack et al. [2012] developed a more accurate model able to capture the 'High Dimensional Dynamic Information Space' [Jack and Schyns 2017]. A very high number of stimuli were created that correspond to sequences of AUs where the temporal course of each AU and its intensity vary. Naive participants evaluated each stimulus in term of emotion labels. Using reverse correlation methods allowed Jack and colleagues to build the mapping between stimuli and the perception of the facial expressions of emotions. Interestingly, the authors found cultural differences in the perception of the stimuli, highlighting that part of the face did not play the same role in the perception of the emotion [Jack et al. 2012].

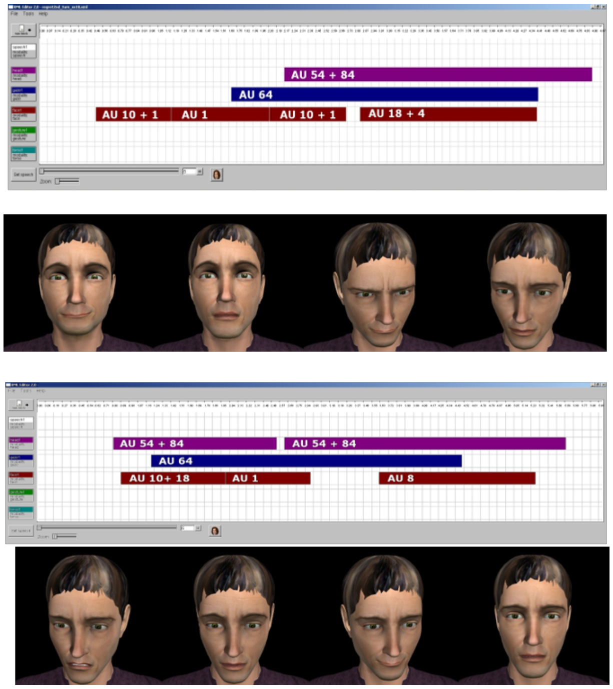


Figure 8.6 Expression of regret as sequences of multimodal behaviors.

Multimodal behaviors Emotions are not expressed solely through facial expressions, but also through the body, voice, touch. Body postures and expressivity contribute to conveying emotions [Pollick et al. 2001, Schindler et al. 2008, Wallbott 1998]. Studies on emotion recognition have been using different stimuli ranging from still images of body poses [Ekman and Friesen 1967] to short videos of actors whose face has been blurred or not, or even videos of actors saying nonsense but phonetically balanced sentences [Bänziger et al. 2006]. While earlier studies reported that the face was more prominent in conveying emotional signals, more recent work found similar results in emotion recognition tasks from face and body stimuli [de Gelder et al. 2015, Kleinsmith and Bianchi-Berthouze 2012]. The first computational models of expressions of emotions for SIAs focused on facial expressions. Later on, several models were proposed for bodily expressions of emotions. To measure the impact of the face and the body, recognition tests were conducted with different stimuli using VAs. Stimuli could be static, an expression at its apex, or short videos [Buisine et al. 2014]. They were defined as displaying expressions of emotion solely through the face, the body and the face and body. In the latter case, the stimuli could be either congruent where the face and body displayed the same emotion, or incongruent where these modalities displayed different emotions [Clavel et al. 2009]. Body expressivity can also characterize the expressions of emotions [Kleinsmith and Bianchi-Berthouze 2012, Wallbott 1998]. It can be defined through different features, such as the amplitude, dynamics or fluidity of the movements [Wallbott 1998] or using dance annotation like Laban [Laban and Lawrence 1974]. Laban Movement Analysis (LMA) schema considers four effort factors: *body*, *effort*, *shape* and *space*. Each of these factors can have various values. For example, *effort* that relates to the quality of the movement is further defined by the parameters: *time*, *flow*, *weight* and *space*. The Expressive MOTion Engine (EMOTE) [Chi et al. 2000] implemented two of the four Laban Movement Analysis factors, *effort* and *space*. This model acts on the movement coordination of the body parts (e.g. relation of the wrist related to the trunk) as well as on the dynamic quality of the movement (such as velocity, acceleration) to compute the expressive motion. Likewise, Hartmann et al. [2005] relied on six motion factors proposed by Wallbott [1998]. These factors were implemented as modifier of the trajectory of an end effector of an articulated chain, such as the wrist. For example, the wrist could change position in space (*spatial* parameter), move in a faster (*temporal* parameter), stronger (*power* parameter), and more fluid (*fluidity* parameter) manner. Later on, this model was extended to also include the *tension* parameter [Huang and Pelachaud 2012].

Encoding - decoding A distinction can also be made related to the encoding or decoding process. An emotion will be expressed, thus encoded, through multimodal behaviors; it will be recognized, thus decoded by the recipients that view these displayed signals. Several studies have highlighted that there may be differences in the encoding and decoding processes that can be due to cultural or individual differences (cf work by Jack and Schyns [2017] just presented and Chapter 13 on culture for SIAs for further details). Scherer and colleagues have applied a modified Brunswikian lens model to explain the different mechanisms involved in the perception and encoding of the expression of emotions [Scherer et al. 2011]. Lhommet and Marsella [2014] report the necessity of designers to take into account such difference when creating expressions of emotions for IVAs.

8.4.2 Rapport

Besides the idea that SIAs should be able to enter in a conversation and perform a well-designed task, the research community started to focus on the social dimensions of an interaction. SIAs should be able to build up a short or long-term relationship with their users. The REA agent [Cassell et al. 1999] mentioned before was given the task of a real-estate agent because in this job it is important to build rapport with the customers through small talk, for instance (see also Chapter 12 on Rapport).

Research on the social aspects of interaction with virtual agents considers many different angles. Various dimensions of social interaction were started to be investigated: rapport, friendship, affiliation, impression management, engagement, or intimacy. On the basis of human-human interactions, Kang et al. [2011] implemented particular patterns of head tilts,

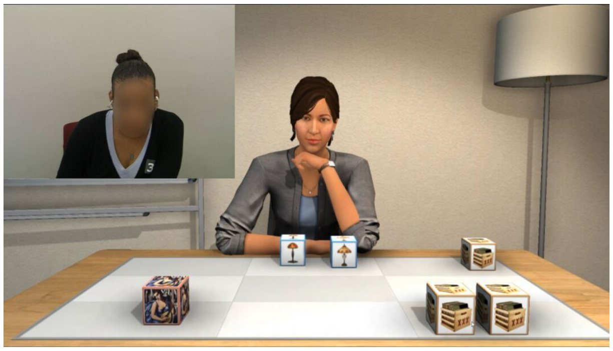


Figure 8.7 Humans are more prone to disclosing information when they believe they are discussing with an autonomous IVA than with an IVA driven by a human operator.

pauses and gaze aversion behaviors in an agent to convey more intimate self-disclosure (see Figure 8.7).

Besides studies into which behaviors best express the desired relational attitude, work on these so-called 'relational agents' has also looked at the appropriate computational models of mind and emotion for such agents. One of the ways in which the computational models were extended was by mechanisms that simulated the concept of 'Theory of Mind' - a person's beliefs about what other persons are feeling and thinking - such as Psychsim [Pynadath and Marsella 2005]. FAtiMA (Fearnot AffecTIve Mind Architecture) [Dias et al. 2014] is another example of offering modeling agent's mental and emotional states. It has been applied to the development of a serious game for children to learn about bullying [Aylett et al. 2007]. Scenarios were designed that involved three main virtual characters, the bully, the victim, and a friend of the victim. The behaviors of the characters were driven by FAtiMA that computed their emotional state and coping behaviors.

The growing attention to relational aspects did not only lead to more studies on the behaviors of agents to convey interpersonal stances such as intimacy and computational models such as PsychSim, but it also led to implementations and studies in which the actions of the agents were more closely aligned and contingent with the verbal and nonverbal behaviors of the human interlocutor. The virtual rapport agent [Gratch et al. 2006, Huang et al. 2011] is based on the idea that contingency of the feedback is more important than the frequency of feedback [Gratch et al. 2007].

One way in which attention to contingent human-agent behavior was given form was through close loop interactions. By paying close attention to detecting the end of turn of a user [de Kok and Heylen 2009] the timing of the agent's response can be manipulated to mark differences in personality and attitude [ter Maat M. 2010]. Another way to have tighter interactions is to have the agent respond when it is listening to the human interlocutor

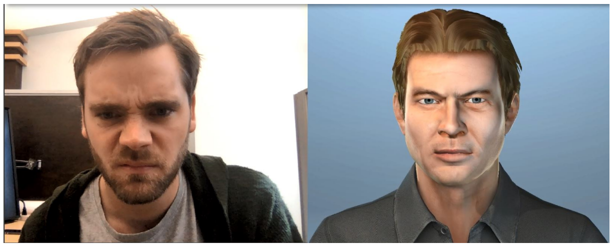


Figure 8.8 IVA imitating human expression during split-steal game.

through backchannels or mimicry (see Figure 8.8). Backchannels can be both verbal ("uhuh, mmm") and nonverbal; a head nod or shake, or a smile in response. To choose the appropriate backchannel and have the SIA perform it at the right moment requires paying close attention to the human interlocutor not only by analyzing the speech and prosody but also the nonverbal expressions in the phase through computer vision.

The SEMAINE project [Schroder et al. 2011] extensively worked on building listener models for agents that would be able to detect appropriate places where an agent should backchannel [de Kok et al. 2013], what strategies to employ [Poppe et al. 2010] and what the effects of variations in backchanneling strategies are on conveying specific relational behaviors or personality [Bevacqua et al. 2012].

With respect to backchannels, Buschmeier and Kopp [2014] have studied the other way: how an agent can elicit feedback from a user depending on the agent's information need (see Figure 8.9).

8.4.3 Personality

People do not behave in the same way. Their personality may affect their decision, intention and also behaviors. As with many complex concepts, there is not yet a definition and representation of personality that receives the consensus of the psychology community [Bergner 2020]. Several models have been proposed to characterize personality along different dimensions. We can name a few models such as OCEAN (Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism), proposed by McCrae and Costa Jr [2008], or the model proposed by Eysenck [2012] (see Chapter 18 on adaptive personality). While there is no consensus on the definition of personality, there is a consensus that behaviors can be linked to different personality traits. Studies have shown how some gestures and postures can be associated to degrees of dominance (from dominant to submissive). Facial expressions, gaze behavior and turn-taking mechanisms are also markers of specific traits. To create SIA with different personalities, computational models have been proposed that act on the type of behaviors and their expressivity the SIAs display. One of the first models was developed by Magnenat-Thalmann and colleagues. Kshirsagar and Magnenat-Thalmann [2002] relied on

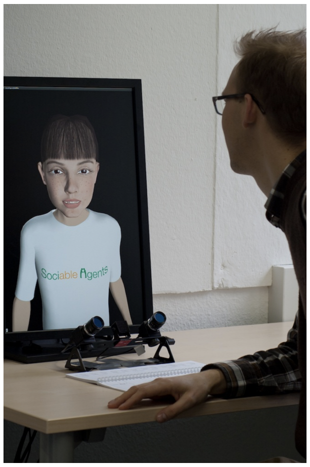


Figure 8.9 IVA eliciting feedback from user [Buschmeier and Kopp 2014] (Copyright CC-BY Hendrik Buschmeier).

the OCEAN representation of personality. The authors link mood, emotion and personality and model the effect of personality trait on the display of an emotion through the introduction of a layer representing the mood of the agent. The links between these three layers, personality, mood and emotions, are represented through a Bayesian Belief Network. From the same lab, Egges et al. [2003] focused on idle movements done while not performing a specific task (i.e. movements done while waiting for a bus). An animation model of posture shifts based on motion capture data allows simulating different idle behaviors.

Other models were proposed for IVAs interacting with a user. In the already mentioned SEMAINE project, four characters were created to act as sensitive artificial listeners [Schroder et al. 2011]. Each character is characterized with a personality trait and a specific emotional agenda [Bevacqua et al. 2012, McRorie et al. 2011]. In this work, personality is defined by three dimensions following Eysenck [2012], namely extraversion-introversion, neuroticism-emotional stability, and psychoticism. Each dimension is associated with different behavior types and the expressivity that were implemented using the notion of baseline and modality preferences. Baseline represents the tendency to perform behaviors [Mancini and Pelachaud 2008]; for example, should they be displayed with large amplitude, and fast and strong movement. This model was applied to IVAs being a speaker and a listener.

8.4.4 Social attitudes

Social attitudes affect how one relates to other members of an interaction. They are often described along two axes, dominance and liking, of the Argyle's Interpersonal Circumplex [Argyle 1988]. Scherer [2005] defines social attitude as "an affective style that spontaneously

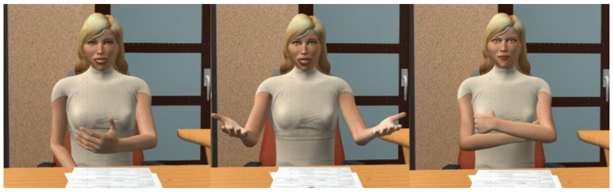


Figure 8.10 Expression of social attitudes using sequence mining.

develops or is strategically employed in the interaction with a person or a group of persons, coloring the interpersonal exchange in that situation". To capture this coloring aspect, Chollet et al. [2017] and then Dermouche and Pelachaud [2016] view the expression of social attitudes as sequences of multimodal behaviors (see Figure 8.10). Both approaches rely on the analysis of a video corpus that was annotated on two main levels: the dimensions of the social attitudes and the multimodal behaviors. The authors aimed to measure which multimodal signals triggered a change in perception of the social attitudes. To this aim, they proposed computational models to extract which common features happen during a change in perception. Sequence mining [Dermouche and Pelachaud 2016] is applied to capture not only the multimodal signals, but also their temporal relationships. These models allow simulating an increase or decrease of either dominance or friendliness, or both, through nonverbal behaviors. Callejas et al. [2014] integrate verbal and nonverbal behaviors in their social attitude model. The authors implemented rules from the PERSONAGE model [Mairesse and Walker 2011] that indicate how some verbal cues such as the use of pronouns, the choice of nouns versus verbs, and the degree of formality, are linked with the expression of personality traits. On the nonverbal side, Callejas et al. [2014] applied a user-centered approach where users selected which gestures shape and expressivity, facial expression, and head nods on an interactive interface would best express a social attitude for a given dialog act. From these data, they built a Bayesian model to select the multimodal behaviors with the highest probability.

8.4.5 Adaptation in interaction

During an interaction, participants adapt to each other. Adaptation may arise at different levels and may take different forms of imitation, synchronization or coordination. Participants may align linguistically, using similar levels of politeness, vocabulary, and grammatical structure [Pickering and Garrod 2004]. They may also imitate a body posture, head movement; they may respond to a smile with a smile, or laughter with laughter [Burgoon et al. 2007]. Adaptation may have several functions in the interaction. It is a strong marker of rapport building, affiliation, engagement and empathy [Lakin and Chartrand 2003]. Its role is very important in social interaction (see Chapter 18 on adaptive personality). Several models have been proposed to enhance the social capabilities of virtual agents. Bailenson and Yee [2005] studied the chameleon effect on social influence. The authors manipulated the behaviors of a virtual agent that would imitate, or not, the head movement of the human interlocutor with delays that could go up to 4 seconds. An agent that imitates their interlocutor's behavior was found more persuasive and more positive than an agent that did not display imitated behaviors.

Biancardi, Dermouche and colleagues have conducted several studies to measure the influence of different adaptation mechanisms [Biancardi et al. 2019b, Dermouche and Pelachaud 2019]. They have developed an architecture where an agent converses with a user while adapting its behavior [Biancardi et al. 2019b, Mancini et al. 2019]. Three adaptation mechanisms were implemented. For two adaptation mechanisms, the agent would adapt dynamically its conversational strategies [Biancardi et al. 2019a] and its multimodal behaviors [Biancardi et al. 2019b] to appear warmer or more competent. Reinforcement learning approaches were used to learn which strategies and which behaviors would optimize either user's engagement or user's impression of the agent. The third adaptation mechanism was obtained at the signal level [Dermouche and Pelachaud 2019]. It took as input the behaviors of both, the user and the agent, over a certain time window, to compute the adapted behavior of the agent. In this last model, a sequence mining model was applied. These three models were evaluated in a science museum following a similar protocol where the agent acted as a guide to a video game exhibit. Results of the studies showed that agents that can adapt to users are perceived either as more competent or warmer. Users had also a better experience of the interaction when interacting with an agent that adapts in some forms.

8.4.6 Social touch

We see IVAs talking and gesturing to us on a screen or we look at them through our virtual reality glasses. We can hear them talk or see them communicate verbal messages through text balloons or other graphics. IVAs can be seen and heard, but their screen-based existence does not seem to allow us to get into physical contact with them. We cannot touch them. They do not touch us. This in contrast to the affordances that physical robots offer.

In human-human communication, touch plays an important role. As we greet each other through a hand shake we establish contact but also show how we appreciate the relation. How long we shake, how much pressure we apply, the temperature of the other hand and how dry or sweaty the hand is convey many impressions. Are we friends?, is this just formal?, are we welcome? and more; these are all questions that might be answered through how we experience the handshake. A handshake might be refused at a greeting or replaced by one, two, three or four kisses on the cheek. Or instead of that, the handshake and kisses a hug might be chosen as the way to greet. Cultural parameters, level of intimacy, how well we are acquainted, the situation: are we meeting friends or meeting for business, as a wedding party or a funeral: these will all be conveyed through whether and how we shake hands, kiss or hug. And then greetings are just one type of interaction in which touch plays a role. An overview of the science of interpersonal touch can be found in Gallace and Spence [2010]. Dibiasi and Gunnoe [2004] is one of the studies on how gender and culture differences play a role in touch. The way touch communicates meaning is discussed in Jones and Yarbrough [1985] and how it communicates emotion is presented in Hertenstein et al. [2009].

The studies on touch in SIAs are almost all motivated by their potential to establish positive effects in the relationship with the human interacting with them. A typical example is the design of the Paro robot [Wada et al. 2006], the robotic seal that responds to a person's touch such as gentle strokes to foster the comfort and feeling of well-being. Also illustrative is the work by Yohanan and MacLean [2012], in which a touch-sensitive robot, the Haptic Creature, is presented. Important for a touch-sensitive robot is that it can recognize a human touch, reason about the meaning and show a reaction. In the paper, they investigate the typical ways in which people try to communicate an emotion through touch and what type of reaction they expect back.

But what about IVAs that are present only through images and sounds? How can we touch a virtual agent? Nguyen et al. [2007] explored the development of a virtual skin for their agent placed in a virtual reality environment. Through motion tracking of the hand of a person, the agent can then feel where a person is touching it and react accordingly. This touch is not complete however, as normally touch is reciprocal and both the toucher and the touched touch each other.

In order for an agent to touch us they need to become hybrids and be extended with a physical device that could conversely also serve as an input device when equipped with touch sensors to touch the IVA (Huisman et al. [2013]).

One of the first studies on extending virtual agents with touching devices was aimed at seeing how a gentle squeeze of a user's hand by an agent using an air bladder could express empathy to users in distress [Bickmore et al. 2010]. The authors pointed out that it is important to recognize the relevance of the different parameters involved in a squeeze such as the pressure, the duration or the number of squeezes. These influence how people perceive the affect arousal of the agent. Another aspect to take into consideration is the interaction of the touch modality with the other verbal and nonverbal communication modes. In their case, they looked at the interaction of touch with facial displays and prosody and found, for instance, that facial displays dominate the perception of affect valence. Also important is the difference between people in their sensitivity to touch.

In the study by Huisman et al. [2014], it was examined how a simulated social touch by an agent would influence the perceived trustworthiness, warmth and politeness of the agent. A participant in the experiment would play either a cooperative or competitive game with two agents in augmented reality. After the game, one of the agents would touch the participant's arm. The agent would approach the participant who would see the agent's arm tapping on the shoulder and simultaneously a tactile display would give a tactile sensation on the participant's arm. The study found no differences between the cooperative and the competitive condition, but the agent touching the participants was observed to score higher on warmth than the non-touching agent.

Both of these studies connected the virtual and the physical environment through physical actuators; an air bladder in the case of Bickmore et al. [2010] and vibration motors in the case of Huisman et al. [2014]. Some effort has been made to simulate touch by thermal feedback Tewell et al. [2017], and force feedback actuators [Bailenson et al. 2007]. These devices are limited in replicating the full dimensions of human touch which involves texture, pressure, temperature and moistness. More involved devices that can simulate human-like touch are being engineered by several researchers in particular as part of robotic devices. An example of this effort is the work by Teyssier et al. [2018, 2020] who among other things attempts to make artificial skin that looks and feels like human skin.

The research on social touch for SIAs is still in its infancy. There are limitations in terms of the technology and there are methodological issues because of the complex factors that are involved in touch. Willemse et al. [2017] reflect on the opportunities and limitations of human-robot interaction. They point out the difficulties in mimicking human-human touch behavior because of the many variables that make it different such as the appearance of the robot, the location where touch is applied, the difference in social context. Touch is a very intimate way to communicate and is only appropriate in particular social contexts governed by cultural and personal practices. It might take some time before robots and agents can take part in these practices.

8.4.7 Machine learning and data driven

An important development in the area of generating nonverbal behaviors is the use of machine learning algorithms to learn the relation between the different modalities, for instance between gestures and speech. These models have been designed and built by relying on an important effort to collect and annotate multimodal databases to directly learn the complexity in the relationship between nonverbal behaviors and speech (See Section 8.5 for discussion on databases). During human interactions, nonverbal behaviors and speech are intrinsically synchronized to convey a message [Cassell et al. 1994]. Hand movements [McNeill 1992], head motion [Heylen 2005, Sadoughi and Busso 2017b] and facial displays [Chovil 1991] play important roles during a conversation, emphasizing a message, clarifying ambiguities, and parsing sentences. As a result, different behaviors are highly correlated not only with speech, but also between themselves [Ding et al. 2013a, Mariooryad and Busso 2012, Sadoughi and Busso 2017a]. These relationships can be learned with machine learning algorithms.

Graphical Models Given the temporal relationship between the various modalities in nonverbal behavior, studies have used different variation of graphical models with explicit connections between frames. An appealing approach to synthesize nonverbal behaviors is hidden Markov models (HMM). Busso et al. [2007] clustered head orientation space, learning models for each of the clusters. The transition between clusters was learned using an HMM formulation. Ding et al. [2013b] explored alternative variations of HMMs to capture the relationship between speech and eyebrow motion. Hofer and Shimodaira [2007] proposed the concept of motion units for head motion, which were generated with HMM. Le et al. [2012] modeled the kinematic feature of head motion with prosodic features using Gaussian mixture models (GMM). Ding et al. [2013a] used a variation of HMMs to model the relation between eyebrow, head movements and speech prosody. The approach incorporates contextual information into the HMM framework. This formulation directly modeled the relationship between head and eyebrow motion.

Other popular graphical models are dynamic Bayesian networks (DBNs). DBNs offer the flexibility to explicitly model connections not only between speech and behaviors, but also across behaviors. For example, Mariooryad and Busso [2012] proposed a speech-driven DBN that explicitly models the relation between head and eyebrow movements. These models also allow us to add explicit discourse function constraints to generate meaningful behaviors [Sadoughi and Busso 2019, Sadoughi et al. 2017] (e.g., head shakes for negations, specific gestures for questions). Another popular graphical model used to synthesize nonverbal behaviors is conditional random fields (CRF). Unlike HMM, which is a generative model with directed graphs, CRF is a discriminative model with undirected graphs. Levine et al. [2010] argued that prosodic features provide valuable information to derive the kinematic of the behavior rather than its shape. They proposed a gesture controller that learns the kinematic of gestures with a speech-driven model based on CRF. Lee and Marsella [2017] considered a formulation based on latent-dynamic conditional random fields (LDCRF). This model incorporates hidden states to capture dynamic within and across gestures. They demonstrated that this approach led to better performance for head nod and eyebrow motion than other models such as HMM and CRF.

Deep learning models Recent studies have relied on deep learning solutions to generate nonverbal behaviors. Ding et al. [2014a, 2015a] proposed to synthesize head movement with fully connected deep neural networks (DNNs), showing that this formulation led to better performance than HMMs [Ding et al. 2014a]. Similar findings were reported by Parker et al. [2017], where they proposed a text-driven model implemented with DNNs to synthesize facial expressions. The approach simultaneously generated facial expressions for different emotions. The results outperformed a baseline implemented with HMMs. Taylor et al. [2016] proposed the sliding window deep neural networks (SW-DNNs) to generate lip motion. The approach synthesized facial parameters representing lip movements driven from speech. The approach generated overlapped visual predictions that are later averaged. The approach was later adapted to synthesize lip movements from phoneme labels [Taylor et al. 2017]. Kucherenko et al. [2019] proposed a deep learning approach based on denoising autoencoder (DAE) to generate hand gestures from speech. The approach creates a bottleneck representation by training an encoder and decoder for hand gesture motions. Then, the approach creates a mapping between speech features and the motion bottleneck representation.

Recurrent Neural Networks Studies have also relied on long-short term memory (LSTM) to capture the temporal relationship in the generation of nonverbal behaviors [Hasegawa et al. 2018]. Fan et al. [2016] proposed a speech-driven approach to synthesize facial parameters associated with the lower facial area. The approach relied on two layers of bidirectional long-short term memory (BLSTM) cells. While BLSTMs are not causal models, they offer improved contextual information by incorporating past and future frames leading to better models. Ding et al. [2015b] demonstrated that BLSTM models led to better performance in generating head movements from speech than fully connected DNNs. They obtained the best performance by combining fully connected layers with BLSTM layers. Similar results were reported by Haag and Shimodaira [2016], who proposed to use the bottleneck layer of one speech-driven network as the input of a second BLSTM-based architecture that predicts head motion. The results were better than a single DNN framework. Sadoughi and Busso [2018b] proposed a multitask learning formulation based on BLSTMs to synthesize nonverbal behaviors. The primary task was generating expressive lip motions. The secondary tasks were recognizing the phonemes and the emotion on the sentence. The LSTM layers create a share feature representation aiming to generate realistic lip motions with the right emotion and properly synchronized with lexical content. Li et al. [2016] also used BLSTM to map speech features with lip movements and facial expressions. Their study explored the generation of articulatory movements related to lexical content (i.e., matching the correct phone) and emotional content (i.e., matching expressions associated with the right emotion). The BLSTM layers capture this relationship driven by acoustic features. Sadoughi and Busso [2017a] proposed a multitask formulation based on BLSTM to create facial movements, where the lower, middle and upper face areas were jointly predicted. This speech-driven formulation considered the interrelationship between facial expressions across different parts of the face. Yunus et al. [2019] use speech-driven recurrent networks with attention mechanism to determine when to generate a nonverbal behavior and its corresponding stroke. The approach focused on beat movements, which are related to the speech rhythm. The addition of attention mechanism was intended to increase the interpretability of the network to understand the relationship between prosodic features and the decision made by the models. Ferstl and McDonnell [2018] proposed an autoencoder implemented with gated recurrent unit (GRU) to model the relation between speech and behaviors. Other methods to generate gestures or facial expressions have explored temporal modeling using convolutional neural networks (CNNs) [Karras et al. 2017], and variational autoencoder (VAE) implemented with BLSTMs [Greenwood et al. 2017a,b].

Generative models A popular generative model that has recently revolutionized several areas is generative adversary networks (GANs). This framework has a generator and a discriminator that are adversarially trained. The generator is trained to create realizations

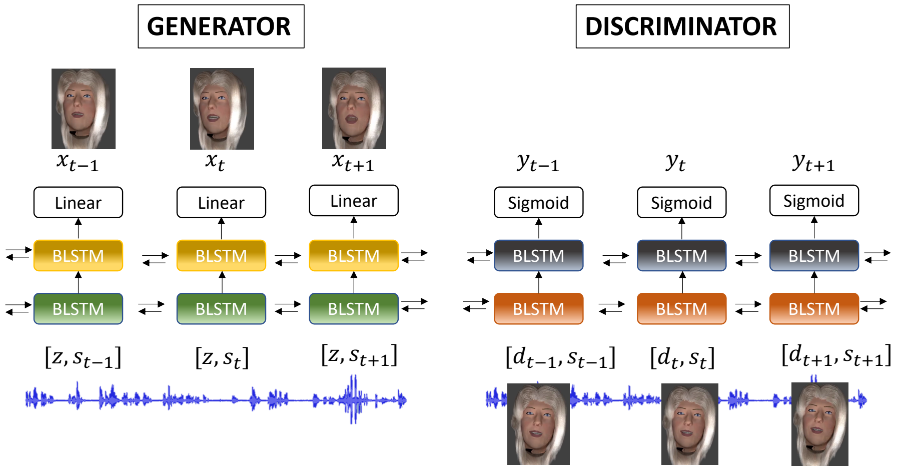


Figure 8.11 Conditional generative adversarial network (GAN) for head movement driven by speech based on Sadoughi and Busso [2018a]. The GAN model was constrained by speech, creating head movements that were temporally synchronized with speech. A similar architecture was also used to synthesize expressive lip movements, where the models were constrained not only by speech, but also by the emotions [Sadoughi and Busso 2020].

that are similar to real examples. The discriminator has to decide whether the input is real or a fake instance created by the generator. After training, the generator is expected to create instances that the discriminator cannot distinguish from real samples. Sadoughi and Busso [2018a] proposed to use conditional generative adversary networks to synthesize head motion, where the discriminator and generator were constrained by the acoustic features (Figure 8.11). The approach was also used to synthesize expressive lip movements [Sadoughi and Busso 2020]. Huang and Khan [2017] proposed the DyadGAN model, which is a two stage framework based on GANs. The first step generates facial sketches using GANs, where the facial expressions of the interlocutor are used to constrain the models. This approach leverages the mutual influence between interlocutors [Mariooryad and Busso 2013]. The second stage also uses GANs to synthesize facial expressions from the sketches. Ferstl et al. [2020] also use generative adversary training to model the relation between speech and nonverbal behaviors. They divided the process into smaller problems, which aim to solve correct behavior dynamics and plausible joint configurations, and to generate diverse and smooth trajectories for nonverbal behaviors. A classifier was used to determine the phase of the behavior, addressing the behavior dynamics. The study found that an adversary loss provides better results than conventional regression losses in mapping the non-deterministic relationship between nonverbal behaviors and speech.

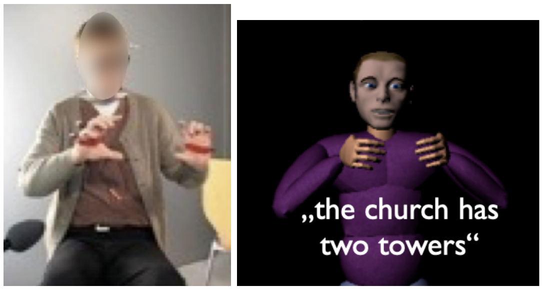


Figure 8.12 (a) motion capture used to gather the GNetIc corpus; (b) Max agent performing an iconic gesture.

8.5 Databases

Computational models of multimodal behaviors have been relying on databases of human data. Over the years, there was a shift in the type of databases that were used. At first models relied on observational studies. Databases were annotated manually as automatic tools were not available yet. Often researchers recorded specific databases needed for their research. Many of these early databases were not made available. Later on, databases were created and gathered larger number of items. They were made accessible to the research community. These databases were mainly used for specific research purposes. We can name the Jaffe [Lyons et al. 1998] and Cohn-Kanade [Lucey et al. 2010] databases. These databases are photos of facial expressions of emotions, mainly the six basic emotions. Their purpose was mainly for facial expression recognition, but also for facial expression synthesis. Later on, video corpus was used. It allows having multimodal synchronized data. The SEMAINE database [McKeown et al. 2011] is one of the few databases that contains videos of human participants interacting with IVAs, the SAL (Sensitive Artificial Listener) agents with different personality traits [McRorie et al. 2011]. Other corpora were gathered for specific research purposes such as the Distress Analysis Interview corpus collected by Gratch et al. [2014] or the negotiation task corpus [Gratch et al. 2016]. The GEneva Multimodal Emotion Portrayals GEMEP database [Bänziger et al. 2006] contains videos of 10 professional actors saying nonsense sentences with 18 affective states. The actors followed induction technique by acting out simple scenarios. Bergmann and Kopp [2009] aimed to study iconic gesture production (see Figure 8.12(b)). To this aim the authors gathered a corpus, called Gesture Net for Iconic Gestures GNetIc, of participants giving direction to another person. Their gestures were captured with video and using a 3D sensor on the hands of the speakers (see Figure 8.12(a)). Rehm et al. [2007] focused on cultural differences in nonverbal behaviors gathered in the Cube-G corpus. To this aim, the authors recorded the Cube-G corpus of dyadic humans interaction in Japan and in Germany.

To have direct access to the signals data and their dynamics, motion capture can be used. Some mocap databases focus on multimodal behavior during an interaction or on body motion quality. For the former, the IEMOCAP recorded by Busso et al. [2008] gathered data of multiple actors interacting in dyads and displaying a variety of emotions. It contains 12h of mocap and video data. The MSP-AVATAR corpus [Sadoughi et al. 2015] contains 6 actors performing improvisation scenarios in dyads. The purpose of this corpus is to study the role of discourse functions in view of modeling SIAs. The CMU Graphics Lab Motion Capture Database[1] contains many examples of actions and behaviors executed with different emotions. On the same line, Fourati and Pelachaud [2016] gathered data of expressive actions movements; each action is performed 3 times by 11 actors with 8 emotions. These last two databases allow, in particular, studying body motion quality.

Lately with the development of models relying on deep learning techniques, the need for huge quantities of data became urgent. Collecting videos or motion capture data in very large quantity require important human resources, infrastructure and can be very time consuming. It is not always feasible in research labs. Thus, many researchers turned their attention to data from the web such as YouTube and TEDx videos. For example, the CMU Multimodal Opinion Sentiment Intensity (CMU-MOSI) [Zadeh et al. 2016] contains a high number videos from movie reviews on YouTube. Reviews are made by 93 participants. It allows having access to a variety of speaker styles. The YouTube Gesture Database [Yoon et al. 2019] is made of TEDx videos. The authors provide tools to segment the videos into scenes that are pertinent for the study (here hand gestures need to be clearly visible in the video segments). Ginosar et al. [2019] aim also to model communicative gestures, paying particular attention to gesture style. The authors gathered 144h of videos from 10 speakers that are made available at the web site[2]. They have also made available the code to extract, train and validate the data.

Collecting data is not an easy task. It is time consuming and it also requires setting the protocol very carefully. Data that are recorded in the lab rely on actors, be they professional or not. It is much more difficult to gather naturalistic data in the lab. To this aim, induction techniques based on scenarios have been used [Bänziger et al. 2006]. Interested readers can read works by Cowie et al. [2011] on issues for collecting data and Cowie et al. [2011], Jokinen and Pelachaud [2013] on issues for annotating data.

8.6 Similarities and Differences in IVAs and SRs

Several studies have explored similarities and differences in virtual and physical agents. 3D animation models for IVAs and mechanical models for SRs have an impact on their behavioral capacities. Moreover the more realistic SIAs are the more expectancies they behave like humans are high. This phenomenon is referred as the uncanny valley [Mori et al. 2012].

[1] http://mocap.cs.cmu.edu/

[2] https://github.com/amirbar/speech2gesture

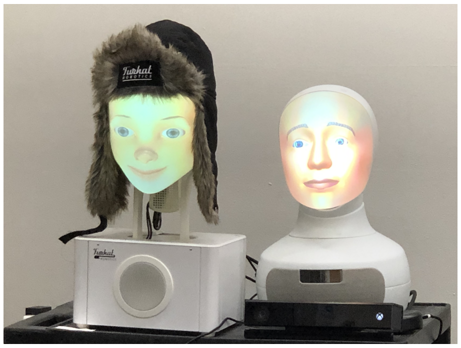


Figure 8.13 First and second generation of furhat robotics.

Studies have been conducted to understand the impact of appearance, the degree of realism in rendering and in behavior on user's perception. The Chapter 4 on Appearance presents in detail studies on rendering styles on a 2D screen and in virtual reality. We can name the works by Rachel McDonnell's group and Bilge Mutlu's group. Some research has been conducted to reach a high degree of realism in physical robots. The work by Hiroshi Ishiguro and his colleagues have achieved amazing results. These researchers do not compare their robots with virtual agents but with humans [MacDorman and Ishiguro 2006]. The Erica robot is another example of human-like robot that has been involved in comparison studies [Inoue et al. 2020]. In contrast, some studies have looked at minimalist representations of agents (e.g. just eyes or a microphone) and checked how behaviors of such minimalist representations can act and be perceived as social agents of an interaction [Tennent et al. 2019]. Other studies, reported in the Introduction Chapter, compared the physical and virtual representation of robots on their impact on task performance and user's perception [Deng et al. 2019]. We can name one line of research that merges both virtual and physical representations. Furhat [Al Moubayed et al. 2012], a robot head with a projected virtual face, is one such example (Figure 8.13). It can display subtle facial expression and gaze in the 3D world. Currently, it is just a head with a face, it does not have body and hand gestures.

Virtual and physical SIAs may share similar decision and emotion models, intention and behavior planners in the SAIBA sense. The previous sections of this chapter present computational models of behaviors that can be implemented in IVAs or in SRs. Their behavior realization varies due to their different capacities in embodiment. However, similarities and differences between social interactive agents with virtual or physical embodiment may arise from other directions as well:

- expressivity capacity that arises from the different degrees of freedom in the face and the body, dynamism of movements, sense of gravity.

- available modalities such as including the possibility to touch and to be touched, to perform facial expression or not.
- a priori of the SIA since people's a priori conception of a virtual agent VA or of a robot SR may affect how they perceive and interact with a SIA; as such the choice of nonverbal behaviors for an SIA may need to take this a priori in consideration.
- spatial and social presence that also affect the perception of personal space with proxemics distance, thus influencing the proxemics and gaze behaviors of SIAs
- exploration capability that allows moving and gazing around; SRs have much greater capability for exploring the surrounding than an IVA displayed on a 2D screen.
- availability depending on the type of displays, be on a large screen, a mobile, or be tiny (to be hold in one hand) or human-size robots. The choice of a behavior could be adapted. Indeed, visibility of a behavior can be impacted when it is being displayed on a large or a tiny screen, for example. Identically, large arm movements executed by human-size robots or by tiny robots will not have the same impact in user's perception.
- degree of realism in virtual and robotic appearance at the level of morphology, rendering of the skin, artifacts (hair, eyelids...). It can have an impact in the perception of the SIAs.

8.7 Current and Future Challenges

In the previous sections we have presented the large advances in generating multimodal behaviors. However much more remains to be done to obtain natural and expressive behaviors for SIAs. We list here a few challenges. This is not an exhaustive list. Other challenges are embedded with modeling emotion, adaptive personality, appearance, long term interaction, and foreseen applications… that are presented in other chapters of this book. It would be highly desirable to tackle all challenges in a holistic manner. However, we will now focus on challenges for generating multimodal expressive behaviors, which represent already difficult tasks to tackle.

Automatic generation of multimodal behaviors The algorithms to generate behaviors need to be designed by considering the intention of the message. The gestures need to respond to the communication function during an interaction. A key challenge is to automatically generate gestures that have meaning and can be easily integrated with other modules of a social interactive agent framework. The most powerful generative models today are adversary networks including variations of GANs. The straightforward approach to constrain the models to specific communication functions is by adding the constraints as inputs of the generator and discriminator. This approach may not scale well when multiple constraints are needed (e.g., hand gestures of a sad SIA while asking a question). As databases have only partial information to train these types of systems, the training strategies of these algorithms will have to consider partial information, leveraging the knowledge extracted from multiple databases.

Databases Another related challenge to create gestures with a clear meaning is to design databases with clear annotations of discourse functions. The MSP-AVATAR corpus [Sadoughi et al. 2015] was collected explicitly for this purpose, providing annotations for a broad range of discourse functions (e.g., contrasting, confirmation, negation, questioning, uncertainty, suggestion, giving orders, and warning). Increasing the number of databases with these labels can serve as a starting point to train deep learning models that are constrained by appropriate discourse functions. Some of the relevant communicative functions needed to generate meaningful nonverbal behaviors can be directly obtained with advances in natural language processing and automatic speech recognition, without requiring manual annotations. This approach can facilitate the collection of large-scale databases. For example, databases collected from video-sharing websites can be very effective in training algorithms for nonverbal behaviors [Vidal et al. 2020, Zadeh et al. 2018].

Social signals Most of the works presented above have focused on communicative gestures, facial expressions, and gaze behaviors. However, there exists a large set of signals that are fully part of the social interactions that have not been, or barely, considered so far. We can name laughter, yawning, cries, hesitation, sighs, etc. These signals can carry a variety of communicative functions and be linked to social stances [Curran et al. 2018, Scott et al. 2014]. Mazzocconi et al. [2020] have proposed a taxonomy highlighting their propositional content. Considering these signals require understanding their communicative functions, and also simulate their animation. Several attempts exist regarding laughter [Ding et al. 2017, El Haddad et al. 2016]. However, these works focused on hilarious laughter only. Modeling laughter animation is quite complex. Laughter involves synchronized torso and head movement, a large variety of facial expressions; breathing and inhalation are also an important part of laughter. Speech laughter is a whole other issue. To our knowledge, few, if not no study has been conducted to study lip movements during speech and laughter at the acoustic level (see Chapters 6 on speech synthesis and 20 on animation). Another issue is to understand where to place a given laughter. How do we respond to laughter? What triggers laughter? Which type of laughter should we produce? Should it carry a specific content? Should it be an answer to interlocutor's behavior? Should it arise from contagion? These are some of the questions to answer to endow agents with laughing capabilities. Producing signals of laughter, sighs, cries, etc. require also taking into account their potential impact on the interaction. One needs to understand how these signals produced by SIAs in a given social context affect the perception of the SIA and of the interaction. Preliminary studies on smiles and hilarious laughter [Ding et al. 2014b, Ochs and Pelachaud 2013] have shown how human users changed their perception of the SIAs depending on the signals they produced and when they displayed them. They also modulate how they perceive what the SIAs are saying.

Go beyond generic model to simulate SIAs with identity Defining what makes an individual is extremely complex. It involves intricate patterns as so many features are interwoven.

We can name some that shape an individual: socio-cultural background, previous personal history, personality traits, emotional tendency, social attitude, interpersonal relationships, competences, knowledge, beliefs, etc. At the behavior level, an individual may have a specific style, have a given behavior expressivity, or display idiosyncratic behaviors, etc. The list of features that makes a person unique could go on. It is huge and very diverse. In this chapter, we have presented several approaches that modeled one or more of these features. These models offer to simulate SIAs with some specificities, but not yet with an identity. In early works, Hayes-Roth and Doyle [1998] created a backstory for the synthetic actors they placed in an application. The backstory would define the role of the agents, their personality profile, but also their family life, hobbies, jobs... The characters would act based on these background information offering motivations for their acting. Noot and Ruttkay [2005] have proposed a representation language, GESTYLE, to capture the behavior "style" of a person. Tags would range from the culture to profession. Dictionaries were created to map these tags into behaviors. These works were a first step to creating agents with identities. However, they tend to create stereotypes. Also they do not model either how the different features characterizing an identity influence each other or what are the processes involved in computing motivations and deciding which actions to perform... Defining SIAs with a mental and affective states [Marsella and Gratch 2009, Pynadath and Marsella 2005] also gives a coherence to SIAs' decisions, actions and emotional states, that are the premises for giving a sense of identity to SIAs. In most cases, SIAs are defined by their role in the interaction. Most of them are young adults. Their cultural background is often not well specified; even though there are studies on modeling cultural agents, this is still not the case for the majority of agents (see Chapter 13 on Culture). SIAs may have a specific appearance but do not correspond to a specific person. One cannot attach an identity to them. Much more research needs to be done to simulate an identity for SIAs that is reflected in their nonverbal behavior.

8.8 Summary and Conclusion

Endowed with virtual or physical eyes, mouths, hands, and arms, SIAs that do not use them to communicate nonverbally in similar ways as humans do raise eyebrows with the human interlocutor. In this chapter we have tried to provide an overview of studies on generating nonverbal expressions of SIAs. We started with introducing concepts from the social sciences and humanities that formed the basis of the computational models and approaches. Interestingly, there has also been a line of research in the social sciences to use VAs to study aspects of how humans communicate nonverbally [de Gelder et al. 2018] [Jack and Schyns 2017].

In the first phase, the behaviors of the SIAs were informed by the studies and theories on nonverbal behavior in humans or on analyzing small corpora of recordings of humans interacting. This led to rule-based systems that made the decision when to perform which nonverbal behavior aligned with speech. Currently, larger corpora are used by machine learning approaches for data-driven generation of nonverbal behaviors.

Implementing nonverbal behaviors for SIAs we need not only pay attention to the quality of realisation in terms of, for instance, expressivity but also to their timing be it in combination with the other expressive modalities of which they make use of such as speech or in close loop interaction with the human interlocutor when providing feedback in the form of backchannels.

Nonverbal communication takes many forms and in the course of 20+ years eye movements, facial expressions, hand, arm gestures, posture have all received elaborate attention with studies on both the repertoire of behaviors that SIAs should be capable of performing and the various functions they serve: from visual prosody such as head nods that are used to emphasize part of the speech to expressions of emotions and interpersonal stance.

This chapter focuses to a large extent on IVAs. For a long time the social robotics community and the virtual agents community went their separate ways. But as more researchers have started to become active in both fields, the studies carried out in one field also become known in the other. In the area of virtual agents, there is a history of collaboration between research labs, resulting in a common language to talk about nonverbal behavior - the Behavior Markup Language - to give one example. With the fields of robotics and agents starting to talk to each other such collaborations will grow and the challenges will be met together.

So, when you start talking to the newest humanoid robot and it does not smile in greeting or none of the other nonverbal behaviors that you expect, it is unlikely that the makers did not read the chapter on Multimodal Behavior in the Handbook on Socially Interactive Agents but more likely that a simple reboot will do the trick.

Bibliography

S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström. 2012. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive behavioural systems*, pp. 114–130. Springer.

I. Albrecht, M. Schröder, J. Haber, and H. Seidel. August 2005. Mixed feelings – expression of non-basic emotions in a muscle-based talking head. *Virtual Reality (Special Issue "Language, Speech and Gesture for VR")*, 8(4): 201–212.

J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio. 2007. The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3-4): 273–287.

E. André, J. Müller, and T. Rist. 1996. The PPP persona: a multipurpose animated presentation agent. In *Proceedings of the workshop on Advanced visual interfaces*, pp. 245–247.

E. André, T. Rist, and J. Mueller. 1998. Integrating reactive and scripted behaviors in a life-like presentation agent. In *Proceedings of the second International Conference on Autonomous Agents*, pp. 261–268.

M. Argyle. 1988. *Bodily Communication*, 2nd. Methuen & Co., London.

O. Arikan and D. Forsyth. July 2002. Interactive motion generation from examples. *ACM Transactions on Graphics (TOC)*, 21(3): 483–490. DOI: 10.1145/566654.566606.

A. Arya and S. DiPaola. 2007. Multispace behavioral model for face-based affective social agents. *EURASIP Journal on Image and Video Processing*, 2007: 1–12.

D. Atkinson, P. Orzechowski, B. Petreca, N. Bianchi-Berthouze, P. Watkins, S. Baurley, S. Padilla, and M. Chantler. 2013. Tactile perceptions of digital textiles: a design research approach. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1669–1678.

J. Austin. 1962. *How to do things with words*. Oxford University Press, London.

R. Aylett, M. Vala, P. Sequeira, and A. Paiva. 2007. Fearnot!–an emergent narrative approach to virtual dramas for anti-bullying education. In *International Conference on Virtual Storytelling*, pp. 202–205. Springer.

J. N. Bailenson and N. Yee. 2005. Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological science*, 16(10): 814–819.

J. N. Bailenson, N. Yee, S. Brave, D. Merget, and D. Koslow. 2007. Virtual interpersonal touch: expressing and recognizing emotions through haptic devices. *Human–Computer Interaction*, 22(3): 325–353.

T. Bänziger, H. Pirker, and K. Scherer. May 2006. GEMEP - Geneva multimodal emotion portrayals: a corpus for the study of multimodal emotional expressions. In *First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation (LREC 2006))*, pp. 15–19. Genoa,Italy.

J. Bavelas, J. Gerwing, and S. Healing. 2014. Hand and facial gestures in conversational interaction. *The Oxford handbook of language and social psychology*, 111: 130.

J. B. Bavelas and N. Chovil. 2000. Visible acts of meaning: An integrated message model of language in face-to-face dialogue. *Journal of Language and social Psychology*, 19(2): 163–194.

K. Bergmann and S. Kopp. 2009. Gnetic–using bayesian decision networks for iconic gesture generation. In *International Workshop on Intelligent Virtual Agents*, pp. 76–89. Springer.

R. M. Bergner. 2020. What is personality? two myths and a definition. *New Ideas in Psychology*, 57: 100759.

J. Beskow. September 1997. Animation of talking agents. In C. Benoit and R. Campbell, eds., *Proceedings of the ESCA Workshop on Audio-Visual Speech Processing*, pp. 149–152. Rhodes, Greece.

J. Beskow and S. McGlashan. August 1997. Olga - a conversational agent with gestures. In *Proceedings of the IJCAI 1997 Workshop on Animated Interface Agents: Making Them Intelligent*. Nagoya, Japan.

E. Bevacqua, E. De Sevin, S. J. Hyniewska, and C. Pelachaud. 2012. A listener model: introducing personality traits. *Journal on Multimodal User Interfaces*, 6(1-2): 27–38.

B. Biancardi, M. Mancini, P. Lerner, and C. Pelachaud. 2019a. Managing an agent's self-presentational strategies during an interaction. *Frontiers in Robotics and AI*, 6: 93.

B. Biancardi, C. Wang, M. Mancini, A. Cafaro, G. Chanel, and C. Pelachaud. 2019b. A computational model for managing impressions of an embodied conversational agent in real-time. In *2019 International Conference on Affective Computing and Intelligent Interaction (ACII)*.

T. W. Bickmore, R. Fernando, L. Ring, and D. Schulman. 2010. Empathic touch by relational agents. *IEEE Transactions on Affective Computing*, 1(1): 60–71.

E. Bozkurt, C. E. Erdem, E. Erzin, T. Erdem, M. Özkan, and A. M. Tekalp. May 2008. Speech-driven automatic facial expression synthesis. In *3DTV Conference 2008: The True Vision - Capture, Transmission and Display of 3D Video*, pp. 273–276. DOI: 10.1109/3DTV.2008.4547861.

M. Brand. August 1999. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques (SIGGRAPH 1999)*, pp. 21–28. Los Angeles, CA, USA. DOI: 10.1145/311535.311537.

C. Bregler, M. Covell, and M. Slaney. August 1997. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques (SIGGRAPH 1997)*, pp. 353–360. Los Angeles, CA,USA.

T. D. Bui. 2004. *Creating Emotions and Facial Expressions for Embodied Agents*. PhD thesis, University of Twente, Departament of Computer Science.

S. Buisine, M. Courgeon, A. Charles, C. Clavel, J.-C. Martin, N. Tan, and O. Grynszpan. 2014. The role of body postures in the recognition of emotions in contextually rich scenarios. *International Journal of Human-Computer Interaction*, 30(1): 52–62.

J. K. Burgoon, L. A. Stern, and L. Dillman. 2007. *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press.

H. Buschmeier and S. Kopp. 2014. When to elicit feedback in dialogue: Towards a model based on the information needs of speakers. In T. W. Bickmore, S. Marsella, and C. L. Sidner, eds., *Intelligent Virtual Agents - 14th International Conference, IVA 2014, Boston, MA, USA, August 27-*

29, 2014. Proceedings, volume 8637 of *Lecture Notes in Computer Science*, pp. 71–80. Springer. https://doi.org/10.1007/978-3-319-09767-1$_$10. DOI: 10.1007/978-3-319-09767-1_10.

C. Busso and S. Narayanan. November 2007. Interrelation between speech and facial gestures in emotional utterances: a single subject study. *IEEE Transactions on Audio, Speech and Language Processing*, 15(8): 2331–2347. DOI: 10.1109/TASL.2007.905145.

C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan. March 2007. Rigid head motion in expressive speech animation: Analysis and synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3): 1075–1086. DOI: 10.1109/TASL.2006.885910.

C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan. December 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42(4): 335–359. DOI: 10.1007/s10579-008-9076-6.

Z. Callejas, B. Ravenet, M. Ochs, and C. Pelachaud. 2014. A computational model of social attitudes for a virtual recruiter. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 93–100.

J. Cassell and M. Stone. 1999. Living hand and mouth. Psychological theories about speech and gestures in interactive dialogue systems. In *AAAI99 Fall Symposium on Psychological Models of Communication in Collaborative Systems*.

J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Bechet, B. Douville, S. Prevost, and M. Stone. 1994. Animated conversation: Rule-based generation of facial expression gesture and spoken intonation for multiple conversational agents. In *Computer Graphics (Proc. of ACM SIGGRAPH'94)*, pp. 413–420. Orlando, FL,USA.

J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. May 1999. Embodiment in conversational interfaces: Rea. In *International Conference on Human Factors in Computing Systems (CHI-99)*, pp. 520–527. Pittsburgh, PA, USA. DOI: 10.1145/302979.303150.

J. Cassell, H. Vilhjálmsson, and T. Bickmore. 2001. BEAT: the Behavior Expression Animation Toolkit. In *Computer Graphics Proceedings, Annual Conference Series*. ACM SIGGRAPH.

D. Chi, M. Costa, L. Zhao, and N. Badler. 2000. The EMOTE model for effort and shape. In K. Akeley, ed., *Siggraph 2000, Computer Graphics Proceedings*, pp. 173–182. ACM Press / ACM SIGGRAPH / Addison Wesley Longman. citeseer.nj.nec.com/costa00emote.html.

M. Chollet, M. Ochs, and C. Pelachaud. 2017. A methodology for the automatic extraction and generation of non-verbal signals sequences conveying interpersonal attitudes. *IEEE Transactions on Affective Computing*.

N. Chovil. 1991. Discourse-oriented facial displays in conversation. *Research on Language & Social Interaction*, 25(1-4): 163–194. DOI: 10.1080/08351819109389361.

C. Clavel, J. Plessier, J.-C. Martin, L. Ach, and B. Morel. 2009. Combining facial and postural expressions of emotions in a virtual character. In *International Workshop on Intelligent Virtual Agents*, pp. 287–300. Springer.

M. M. Cohen and D. W. Massaro. 1993. Modeling coarticulation in synthetic visual speech. In M. Magnenat-Thalmann and D. Thalmann, eds., *Models and Techniques in Computer Animation*, pp. 139–156. Springer-Verlag, Tokyo.

M. Courgeon, J.-C. Martin, and C. Jacquemin. 2008. User's gestural exploration of different virtual agents' expressive profiles. In *Proceedings of the 7th international joint conference on Autonomous* *agents and multiagent systems-Volume 3*, pp. 1237–1240. International Foundation for Autonomous Agents and Multiagent Systems.

M. Courgeon, C. Céline, and J.-C. Martin. 2014. Modeling Facial Signs of Appraisal During Interaction: Impact on Users' Perception and Behavior. *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, pp. 765–772. http://dl.acm.org/citation.cfm?id=2615731.2615855.

R. Cowie, E. Douglas-Cowie, M. McRorie, I. Sneddon, L. Devillers, and N. Amir. 2011. Issues in data collection. In *Emotion-Oriented Systems*, pp. 197–212. Springer.

W. Curran, G. J. McKeown, M. Rychlowska, E. André, J. Wagner, and F. Lingenfelser. 2018. Social context disambiguates the interpretation of laughter. *Frontiers in Psychology*, 8: 2342.

N. Dael, M. Mortillaro, and K. R. Scherer. 2012. The body action and posture coding system (bap): Development and reliability. *Journal of Nonverbal Behavior*, 36(2): 97–121.

B. de Gelder, A. De Borst, and R. Watson. 2015. The perception of emotion in body expressions. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(2): 149–158.

B. de Gelder, J. Kätsyri, and A. W. de Borst. 2018. Virtual reality and the new psychophysics. *British Journal of Psychology*, 109: 421–426.

I. de Kok and D. Heylen. 2009. Multimodal end-of-turn prediction in multi-party meetings. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, ICMI-MLMI '09, p. 91–98. Association for Computing Machinery, New York, NY, USA. ISBN 9781605587721. https://doi.org/10.1145/1647314.1647332. DOI: 10.1145/1647314.1647332.

I. de Kok, D. Heylen, and L. Morency. 2013. Speaker-adaptive multimodal prediction model for listener responses. In J. Epps, F. Chen, S. L. Oviatt, K. Mase, A. Sears, K. Jokinen, and B. W. Schuller, eds., *2013 International Conference on Multimodal Interaction, ICMI '13, Sydney, NSW, Australia, December 9-13, 2013*, pp. 51–58. ACM. https://doi.org/10.1145/2522848.2522866. DOI: 10.1145/2522848.2522866.

E. Deng, B. Mutlu, M. J. Mataric, et al. 2019. Embodiment in socially interactive robots. *Foundations and Trends® in Robotics*, 7(4): 251–356.

S. Dermouche and C. Pelachaud. 2016. Sequence-based multimodal behavior modeling for social agents. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 29–36.

S. Dermouche and C. Pelachaud. 2019. Generative Model of Agent's Behaviors in Human-Agent Interaction. In *Proceedings of the 21st ACM International Conference on Multimodal Interaction (ICMI 2019)*. ACM, Suzhou, Jiangsu, China.

J. Dias, S. Mascarenhas, and A. Paiva. 2014. FAtiMA modular: Towards an agent architecture with a generic appraisal framework. In *Emotion Modeling*, pp. 44–56. Springer.

R. Dibiasi and J. Gunnoe. 2004. Gender and culture differences in touching behavior. *The Journal of Social Psychology 144*, 1: 49–62.

C. Ding, P. Zhu, L. Xie, D. Jiang, and Z. Fu. September 2014a. Speech-driven head motion synthesis using neural networks. In *Interspeech 2014*, pp. 2303–2307. Singapore.

C. Ding, L. Xie, and P. Zhu. 2015a. Head motion synthesis from speech using deep neural networks. *Multimedia Tools and Applications*, 74(22): 9871–9888.

C. Ding, P. Zhu, and L. Xie. September 2015b. BLSTM neural networks for speech driven head motion synthesis. In *Interspeech 2015*, pp. 3345–3349. Dresden, Germany.

Y. Ding, C. Pelachaud, and T. Artieres. August 2013a. Modeling multimodal behaviors from speech prosody. In R. Aylett, B. Krenn, C. Pelachaud, and H. Shimodaira, eds., *International Conference on Intelligent Virtual Agents (IVA 2013)*, volume 8108 of *Lecture Notes in Computer Science*, pp. 198–207. Springer Berlin Heidelberg, Edinburgh, UK. ISBN 978-3-642-40415-3. DOI: 10.1007/978-3-642-40415-3_19.

Y. Ding, M. Radenen, T. Artières, and C. Pelachaud. May 2013b. Speech-driven eyebrow motion synthesis with contextual markovian models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, pp. 3756–3760. Vancouver, BC, Canada. DOI: 10.1109/ICASSP.2013.6638360.

Y. Ding, K. Prepin, J. Huang, C. Pelachaud, and T. Artières. 2014b. Laughter animation synthesis. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 773–780.

Y. Ding, J. Huang, and C. Pelachaud. 2017. Audio-driven laughter behavior controller. *IEEE Transactions on Affective Computing*, 8(4): 546–558.

A. Egges, S. Kshirsagar, and N. Magnenat-Thalmann. 2003. A model for personality and emotion simulation. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 453–461. Springer.

P. Ekman. 2004. Emotional and conversational nonverbal signals. In *Language, knowledge, and representation*, pp. 39–50. Springer.

P. Ekman and W. Friesen. 1975. *Unmasking the Face: A guide to recognizing emotions from facial clues*. Prentice-Hall, Inc.

P. Ekman and W. V. Friesen. 1967. Head and body cues in the judgment of emotion: A reformulation. *Perceptual and motor skills*, 24(3 PT 1): 711–724.

P. Ekman, W. Friesen, and J. Hager. 2002. *Facial action coding system (FACS). A human face*. Research Nexus, Salt Lake City.

K. El Haddad, H. Çakmak, E. Gilmartin, S. Dupont, and T. Dutoit. 2016. Towards a listening agent: a system generating audiovisual laughs and smiles to show interest. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 248–255.

H. J. Eysenck. 2012. *A model for personality*. Springer Science & Business Media.

B. Fan, L. Xie, S. Yang, L. Wang, and F. K. Soong. May 2016. A deep bidirectional LSTM approach for video-realistic talking head. *Multimedia Tools and Applications*, 75(9): 5287–5309. DOI: 10.1007/s11042-015-2944-3.

Y. Ferstl and R. McDonnell. November 2018. Investigating the use of recurrent motion modelling for speech gesture generation. In *Intelligent Virtual Agents (IVA 2018)*, pp. 93–98. Sydney, NSW, Australia. DOI: 10.1145/3267851.3267898.

Y. Ferstl, M. Neff, and R. McDonnell. June 2020. Adversarial gesture generation with realistic gesture phasing. *Computers & Graphics*, 89: 117–130. DOI: 10.1016/j.cag.2020.04.007.

N. Fourati and C. Pelachaud. 2014. Collection and characterization of emotional body behaviors. In *Proceedings of the 2014 International Workshop on Movement and Computing*, pp. 49–54.

N. Fourati and C. Pelachaud. 2016. Perception of emotions and body movement in the emilya database. *IEEE Transactions on Affective Computing*, 9(1): 90–101.

W. V. Friesen, P. Ekman, et al. 1983. EMFACS-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*, 2(36): 1.

A. Fukayama, T. Ohno, N. Mukawa, M. Sawaki, and N. Hagita. 2002. Messages embedded in gaze of interface agents — impression management with agent's gaze. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, p. 41–48. Association for Computing Machinery, New York, NY, USA. ISBN 1581134533. https://doi.org/10.1145/503376.503385. DOI: 10.1145/503376.503385.

A. Gallace and C. Spence. 2010. The science of interpersonal touch: An overview. *Neuroscience and Biobehavioral Reviews*, 34(2): 246–259.

S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik. 2019. Learning individual styles of conversational gesture. In *Computer Vision and Pattern Recognition (CVPR)*.

H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang. May 2002. Visual prosody: Facial movements accompanying speech. In *Proc. of IEEE International Conference on Automatic Faces and Gesture Recognition*, pp. 396–401. Washington, D.C., USA.

J. Gratch, A. Okhmatovskaia, F. Lamothe, S. Marsella, M. Morales, R. van der Werf, and L. Morency. August 2006. Virtual rapport. In J. Gratch, M. Young, R. Aylett, D. Ballin, and P. Olivier, eds., *International Conference on Intelligent Virtual Agents (IVA 2006)*, volume 4133 of *Lecture Notes in Computer Science*, pp. 14–27. Springer-Verlag Berlin Heidelberg, Marina Del Rey, CA, USA. ISBN 978-3-540-37593-7.

J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy. 2007. Creating rapport with virtual agents. In *Intelligent Virtual Agents*. Springer, Berlin, Heidelberg.

J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, pp. 3123–3128.

J. Gratch, D. DeVault, and G. Lucas. 2016. The benefits of virtual humans for teaching negotiation. In *International Conference on Intelligent Virtual Agents*, pp. 283–294. Springer.

D. Greenwood, S. Laycock, and I. Matthews. August 2017a. Predicting head pose in dyadic conversation. In J. Beskow, C. Peters, G. Castellano, C. O'Sullivan, I. Leite, and S. Kopp, eds., *International Conference on Intelligent Virtual Agents (IVA 2017)*, volume 10498 of *Lecture Notes in Computer Science*, pp. 160–169. Springer Berlin Heidelberg, Stockholm, Sweden. ISBN 978-3-319-67400-1. DOI: 10.1007/978-3-319-67401-8_18.

D. Greenwood, S. Laycock, and I. Matthews. August 2017b. Predicting head pose from speech with a conditional variational autoencoder. In *Interspeech 2017*, pp. 3991–3995. Stockholm, Sweden. DOI: 10.21437/Interspeech.2017-894.

K. Haag and H. Shimodaira. September 2016. Bidirectional LSTM networks employing stacked bottleneck features for expressive speech-driven head motion synthesis. In D. Traum, W. Swartout, P. Khooshabeh, S. Kopp, S. Scherer, and A. Leuski, eds., *International Conference on Intelligent Virtual Agents (IVA 2016)*, volume 10011 of *Lecture Notes in Computer Science*, pp. 198–207. Springer Berlin Heidelberg, Los Angeles, CA, USA. ISBN 978-3-319-47664-3. DOI: 10.1007/978-3-319-47665-0_18.

M. Halliday. 1967. *Intonation and Grammar in British English*. Mouton, The Hague.

B. Hartmann, M. Mancini, and C. Pelachaud. 2005. Implementing expressive gesture synthesis for embodied conversational agents. In *International Gesture Workshop*, pp. 188–199. Springer.

D. Hasegawa, N. Kaneko, S. Shirakawa, H. Sakuta, and K. Sumi. November 2018. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *Intelligent Virtual Agents (IVA 2018)*, pp. 79–86. Sydney, NSW, Australia. DOI: 10.1145/3267851.3267878.

B. Hayes-Roth and P. Doyle. 1998. Animate characters. *Autonomous agents and multi-agent systems*, 1(2): 195–230.

M. J. Hertenstein, R. Holmes, M. McCullough, and D. Keltner. 2009. The communication of emotion via touch. *Emotion*, 9(4): 566.

D. Heylen. April 2005. Challenges ahead head movements and other social acts in conversation. In *Artificial Intelligence and Simulation of Behaviour (AISB 2005), Social Presence Cues for Virtual Humanoids Symposium*, p. 8. Hertfordshire, United Kingdom.

D. Heylen. 2010. Ubiquitous gaze: using gaze at the interface. In *Human-centric interfaces for ambient intelligence*, pp. 49–70. Elsevier.

D. Heylen, I. van Es, A. Nijholt, and B. van Dijk. 2005. Controlling the gaze of conversational agents. In J. van Kuppevelt, L. Dybkjær, and N. Bernsen, eds., *Advances in Natural Multimodal Dialogue Systems*, volume 30 of *Text, Speech and Language Technology*, pp. 245–262. Springer, Dordrecht. ISBN 978-1-4020-3933-1. DOI: 10.1007/1-4020-3933-6_11.

D. Heylen, S. Kopp, S. C. Marsella, C. Pelachaud, and H. Vilhjálmsson. 2008. The next step towards a function markup language. In *Intelligent Virtual Agents*, pp. 270–280. Springer.

G. Hofer and H. Shimodaira. August 2007. Automatic head motion prediction from speech data. In *Interspeech 2007*, pp. 758–761. Antwerp, Belgium.

J. Huang and C. Pelachaud. 2012. Expressive body animation pipeline for virtual agent. In *proceedings of 12th International Conference of Intelligent Virtual Agents - IVA*, pp. 355–362.

L. Huang, L. Morency, and J. Gratch. September 2011. Virtual rapport 2.0. In H. Vilhjálmsson, S. Kopp, S. Marsella, and K. Thórisson, eds., *intelligent virtual agents*, volume 6895 of *Lecture Notes in Computer Science*, pp. 68–79. Springer Berlin Heidelberg, Reykjavik, Iceland. ISBN 978-3-642-23974-8. DOI: 10.1007/978-3-642-23974-8_8.

Y. Huang and S. M. Khan. July 2017. DyadGAN: Generating facial expressions in dyadic interactions. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2017)*, pp. 2259–2266. Honolulu, HI, USA. DOI: 10.1109/CVPRW.2017.280.

E. M. Huis In‘t Veld, G. J. van Boxtel, and B. de Gelder. 2014. The body action coding system ii: muscle activations during the perception and expression of emotion. *Frontiers in behavioral neuroscience*, 8: 330.

G. Huisman, A. Frederiks, B. V. Dijk, D. Hevlen, and B. Kröse. 2013. The TaSST: Tactile sleeve for social touch. In *World Haptics Conference (WHC)*, pp. 211–216.

G. Huisman, J. Kolkmeier, and D. Heylen. 2014. With us or against us: Simulated social touch by virtual agents in a cooperative or competitive setting. In T. W. Bickmore, S. Marsella, and C. L., eds., *Intelligent Virtual Agents - 14th International Conference, IVA 2014, Boston, MA, USA, August 27-*

29, 2014. Proceedings, volume 8637 of *Lecture Notes in Computer Science*, pp. 204–213. Springer. https://doi.org/10.1007/978-3-319-09767-1$_$25. DOI: 10.1007/978-3-319-09767-1_25.

K. Inoue, D. Lala, K. Yamamoto, S. Nakamura, K. Takanashi, and T. Kawahara. 2020. An attentive listening system with android erica: Comparison of autonomous and WOZ interactions. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 118–127.

R. E. Jack and P. G. Schyns. 2017. Toward a social psychophysics of face communication. *Annual review of psychology*, 68: 269–297.

R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns. 2012. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19): 7241–7244.

K. Jokinen and C. Pelachaud. 2013. From annotation to multimodal behaviour. In *Coverbal Synchrony in Human-Machine Interaction*, pp. 203–222. CRC Press, Taylor & Francis Group.

S. E. Jones and A. E. Yarbrough. 1985. A naturalistic study of the meanings of touch. *Communications Monographs*, 52.

S.-H. Kang, C. Sidner, J. Gratch, R. Artstein, L. Huang, and L.-P. Morency. 09 2011. Modeling nonverbal behavior of a virtual counselor during intimate self-disclosure. In *International Workshop on Intelligent Virtual Agents*, volume 6895, pp. 455–457. DOI: 10.1007/978-3-642-23974-8_60.

T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen. July 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4): 94. DOI: 10.1145/3072959.3073658.

A. Kendon. 1980. Gesticulation and speech: Two aspects of the process of utterance. In M.R.Key, ed., *The Relation between Verbal and Nonverbal Communication*, pp. 207–227. Mouton.

S. Kettebekov, M. Yeasin, and R. Sharma. April 2005. Prosody based audiovisual coanalysis for coverbal gesture recognition. *IEEE Transactions on Multimedia*, 7(2): 234–242.

M. Kipp, M. Neff, K. Kipp, and I. Albrecht. September 2007. Towards natural gesture synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis. In C. Pelachaud, J. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, eds., *International Workshop on Intelligent Virtual Agents (IVA2007)*, volume 4722 of *Lecture Notes in Computer Science*, pp. 15–28. Springer Berlin Heidelberg, Paris, France. ISBN 978-3-540-74997-4. DOI: 10.1007/978-3-540-74997-4_2.

A. Kleinsmith and N. Bianchi-Berthouze. 2012. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4(1): 15–33.

S. Kopp and I. Wachsmuth. June 2002. Model-based animation of co-verbal gesture. In *Proceedings of Computer Animation*, pp. 252–257. Geneva, Switzerland.

S. Kopp and I. Wachsmuth. March 2004. Synthesizing multimodal utterances for conversational agents. *Computer animation & virtual worlds*, 15(1): 39–52. DOI: 10.1002/cav.6.

S. Kopp, B. Krenn, S. Marsella, A. Marshall, C. Pelachaud, H.Pirker, K. Thorisson, and H. Vilhjálmsson. 2006. Towards a common framework for multimodal generation: the behavior markup language. In *Intelligent Virtual Agents - IVA*, pp. 205–217.

S. Kshirsagar and N. Magnenat-Thalmann. June 2002. A multilayer personality model. In *International symposium on Smart graphics (SMARTGRAPH 2002)*, pp. 107–115. Hawthorne, NY, USA. DOI: 10.1145/569005.569021.

T. Kucherenko, D. Hasegawa, G. Henter, N. Kaneko, and H. Kjellström. July 2019. Analyzing input and output representations for speech-driven gesture generation. In *International Conference on Intelligent Virtual Agents (IVA 2019)*, pp. 97–104. Paris, France. DOI: 10.1145/3308532.3329472.

R. Laban and F. C. Lawrence. 1974. *Effort: Economy in body movement*. Plays, Inc., Boston.

J. L. Lakin and T. L. Chartrand. 2003. Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological science*, 14(4): 334–339.

B. H. Le, X. Ma, and Z. Deng. November 2012. Live speech driven head-and-eye motion generators. *IEEE Transactions on Visualization and Computer Graphics*, 18(11): 1902–1914. DOI: 10.1109/TVCG.2012.74.

J. Lee and S. Marsella. 2006. Nonverbal behavior generator for embodied conversational agents. In *International Workshop on Intelligent Virtual Agents*, pp. 243–255. Springer.

J. Lee and S. Marsella. September 2017. Modeling speaker behavior: A comparison of two approaches. In Y. Nakano, M. Neff, A. Paiva, and M. Walker, eds., *International Conference on Intelligent Virtual Agents (IVA 2012)*, volume 7502 of *Lecture Notes in Computer Science*, pp. 160–169. Springer Berlin Heidelberg, Santa Cruz, CA, USA. ISBN 978-3-642-33197-8. DOI: 10.1007/978-3-642-33197-8_17.

J. Lee, J. Chai, P. Reitsma, J. Hodgins, and N. Pollard. July 2002. Interactive control of avatars animated with human motion data. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques (SIGGRAPH 2002)*, pp. 491–500. San Antonio, Texas, USA. DOI: 10.1145/566570.566607.

J. Lester, S. Towns, C. Callaway, J. Voerman, and P. Fitzgerald. 2000. Deictic and emotive communication in animated pedagogical agents. In S. P. J. Cassell, J. Sullivan and E. Churchill, eds., *Embodied Conversational Characters*, pp. 123–154. MITpress, Cambridge, MA.

S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun. July 2010. Gesture controllers. *ACM Transactions on Graphics*, 29(4): 124:1–124:11. DOI: 10.1145/1778765.1778861.

M. Lhommet and S. C. Marsella. 2014. Expressing emotion through posture. *The Oxford handbook of affective computing*, 273.

X. Li, Z. Wu, H. Meng, J. Jia, X. Lou, and L. Cai. September 2016. Expressive speech driven talking avatar synthesis with DBLSTM using limited amount of emotional bimodal data. In *Interspeech 2016*, pp. 1477–1481. San Francisco, CA, USA. DOI: 10.21437/Interspeech.2016-364.

P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pp. 94–101. IEEE.

M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. 1998. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pp. 200–205. IEEE.

K. F. MacDorman and H. Ishiguro. 2006. The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*, 7(3): 297–337.

F. Mairesse and M. A. Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37(3): 455–488.

L. Malatesta, A. Raouzaiou, K. Karpouzis, and S. Kollias. 2009. Towards modeling embodied conversational agent character profiles using appraisal theory predictions in expression synthesis. *Applied Intelligence*, 30(1): 58–64.

M. Mancini and C. Pelachaud. 2008. Distinctiveness in multimodal behaviors. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, pp. 159–166. Citeseer.

M. Mancini, B. Biancardi, S. Dermouche, P. Lerner, and C. Pelachaud. 2019. An architecture for agent's impression management based on user's engagement. In *International Conference on Intelligent Virtual Agents*. Springer.

S. Mariooryad and C. Busso. October 2012. Generating human-like behaviors using joint, speech-driven models for conversational agents. *IEEE Transactions on Audio, Speech and Language Processing*, 20(8): 2329–2340. DOI: 10.1109/TASL.2012.2201476.

S. Mariooryad and C. Busso. April-June 2013. Exploring cross-modality affective reactions for audiovisual emotion recognition. *IEEE Transactions on Affective Computing*, 4(2): 183–196. DOI: 10.1109/T-AFFC.2013.11.

S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro. July 2013. Virtual character performance from speech. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA 2013)*, pp. 25–35. Anaheim, CA, USA. DOI: 10.1145/2485895.2485900.

S. C. Marsella and J. Gratch. 2009. Ema: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1): 70–90.

C. Mazzocconi, Y. Tian, and J. Ginzburg. 2020. What's your laughter doing there? a taxonomy of the pragmatic functions of laughter. *IEEE Transactions on Affective Computing*.

R. R. McCrae and P. T. Costa Jr. 2008. The five-factor theory of personality. In L. P. O.P. John, R.W. Robins, ed., *Handbook of Personality: Theory and Research*, pp. 159–181. Guilford Press.

G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. 2011. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1): 5–17.

D. McNeill. 1992. *Hand and Mind: What gestures reveal about thought*. The University of Chicago Press, Chicago, IL, USA. ISBN 0-226-56132-1.

M. McRorie, I. Sneddon, G. McKeown, E. Bevacqua, E. de Sevin, and C. Pelachaud. 2011. Evaluation of four designed virtual agent personalities. *IEEE Transactions on Affective Computing*, 3(3): 311–322.

A. Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology: Developmental, Learning, Personality, Social*, 14: 261–292.

M. Mori, K. F. MacDorman, and N. Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2): 98–100.

M. Neff, M. Kipp, I. Albrecht, and H. Seidel. March 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics (TOG)*, 27(1): 1–24. DOI: 10.1145/1330511.1330516.

N. Nguyen, I. Wachsmuth, and S. Kopp. 2007. Touch perception and emotional appraisal for a virtual agent. In *Proceedings Workshop Emotion and Computing- Current Research and Future Impact, KI*,

p. 17–22.

R. Niewiadomski and C. Pelachaud. 2007. Fuzzy similarity of facial expressions of embodied agents. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, eds., *Proceedings of the 7th International Conference on Intelligent Virtual Agents (IVA)*, pp. 86–98. Springer.

R. Niewiadomski, S. J. Hyniewska, and C. Pelachaud. 2011. Constraint-based model for synthesis of multimodal sequential expressions of emotions. *IEEE Transactions on Affective Computing*, 2(3): 134–146.

H. Noot and Z. Ruttkay. 2005. Variations in gesturing and speech by gestyle. *Int. J. Hum.-Comput. Stud.*, 62(2): 211–229. ISSN 1071-5819.

M. Ochs and C. Pelachaud. 2013. Socially aware virtual characters: The social signal of smiles. *IEEE Signal Processing Magazine*, 30(2): 128–132.

H. Oster. 2006. Baby facs: Facial action coding system for infants and young children. *Unpublished monograph and coding manual. New York University*.

J. Ostermann. 2002. Face animation in mpeg-4. In I. Pandzic and R. Forchheimer, eds., *MPEG-4 Facial Animation - The Standard Implementation and Applications*, pp. 17–55. Wiley, England.

M. Paleari and C. Lisetti. 2006. Psychologically grounded avatars expressions. In *First Workshop on Emotion and Computing at KI 2006, 29th Annual Conference on Artificial Intelligence*. Bremen, Germany.

X. Pan, M. Gillies, T. M. Sezgin, and C. Loscos. 2007. Expressing complex mental states through facial expressions. In *Second International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 745–746. Springer.

F. Parke. 1975. A model for human faces that allows speech synchronized animation. *Computer and Graphics, Pergamon Press*, pp. 3–4.

F. I. Parke. 1972. Computer generated animation of faces. In *Proceedings of the ACM annual conference-Volume 1*, pp. 451–457.

J. Parker, R. Maia, Y. Stylianou, and R. Cipolla. March 2017. Expressive visual text to speech and expression adaptation using deep neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, pp. 4920–4924. New Orleans, LA, USA. DOI: 10.1109/ICASSP.2017.7953092.

C. Pelachaud and I. Poggi. 1998. Multimodal communication between synthetic agents. In *Advanced Visual Interface*. Aquila, Italy.

C. Pelachaud, N. Badler, and M. Steedman. January-March 1996. Generating facial expressions for speech. *Cognitive Science*, 20(1): 1–46.

M. J. Pickering and S. Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2): 169–190.

I. Poggi. 2007. *Mind, Hands, Face and Body. A Goal and Belief View of Multimodal Communication*, volume Körper, Zeichen, Kultur, (19). Weidler Verlag.

I. Poggi and C. Pelachaud. 2000. Performative facial expressions in animated faces. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, eds., *Embodied Conversational Agents*, pp. 154–188. MIT Press, Cambridge, MA, USA.

F. Pollick, H. Paterson, A. Bruderlin, and A. Sanford. 2001. Perceiving affect from arm movement. *Cognition*, 82: 51 – 61.

R. Poppe, K. P. Truong, D. Reidsma, and D. Heylen. 2010. Backchannel strategies for artificial listeners. In J. M. Allbeck, N. I. Badler, T. W. Bickmore, C. Pelachaud, and A. Safonova, eds., *Intelligent Virtual Agents, 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings*, volume 6356 of *Lecture Notes in Computer Science*, pp. 146–158. Springer. https://doi.org/10.1007/978-3-642-15892-6$_$16. DOI: 10.1007/978-3-642-15892-6_16.

D. V. Pynadath and S. C. Marsella. 2005. Psychsim: Modeling theory of mind with decision-theoretic agents. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, p. 1181–1186. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

M. Rehm and E. André. 2005. Catch me if you can - exploring lying agents in social settings. In F. Dignum, V. Dignum, S. Koenig, S. Kraus, M. P. Singh, and M. Wooldridge, eds., *Proceedings of International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pp. 937–944. ACM, Utrecht, The Netherlands.

M. Rehm, E. André, N. Bee, B. Endrass, M. Wissner, Y. Nakano, T. Nishida, and H. Huang. 2007. The cube-g approach–coaching culture-specific nonverbal behavior by virtual agents. *Organizing and learning through gaming and simulation: proceedings of Isaga*, p. 313.

J. Rickel and W. Johnson. 1999. Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied Artificial Intelligence*, 13: 343–382.

A. Rizzo, U. Neumann, R. Enciso, D. Fidaleo, and J. Noh. July 2004. Performance-driven facial animation: Basic research on human judgments of emotional state in facial avatars. *CyberPsychology & Behavior*, 4(4): 471–487. DOI: 10.1089/109493101750527033.

Z. Ruttkay, H. Noot, and P. T. Hagen. 2003. Emotion disc and emotion squares: tools to explore the facial expression face. *Computer Graphics Forum*, 22(1): 49–53.

N. Sadoughi and C. Busso. November 2015. Retrieving target gestures toward speech driven animation with meaningful behaviors. In *International conference on Multimodal interaction (ICMI 2015)*, pp. 115–122. Seattle, WA, USA. DOI: 10.1145/2818346.2820750.

N. Sadoughi and C. Busso. August 2017a. Joint learning of speech-driven facial motion with bidirectional long-short term memory. In J. Beskow, C. Peters, G. Castellano, C. O'Sullivan, I. Leite, and S. Kopp, eds., *International Conference on Intelligent Virtual Agents (IVA 2017)*, volume 10498 of *Lecture Notes in Computer Science*, pp. 389–402. Springer Berlin Heidelberg, Stockholm, Sweden. ISBN 978-3-319-67400-1. DOI: 10.1007/978-3-319-67401-8_49.

N. Sadoughi and C. Busso. January 2017b. Head motion generation. In B. Müller, S. Wolf, G.-P. Brueggemann, Z. Deng, A. McIntosh, F. Miller, and W. Scott Selbie, eds., *Handbook of Human Motion*, pp. 1–25. Springer International Publishing. ISBN 978-3-319-30808-1. DOI: 10.1007/978-3-319-30808-1_4-1.

N. Sadoughi and C. Busso. April 2018a. Novel realizations of speech-driven head movements with generative adversarial networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, pp. 6169–6173. Calgary, AB, Canada. DOI: 10.1109/ICASSP.2018.8461967.

N. Sadoughi and C. Busso. May 2018b. Expressive speech-driven lip movements with multitask learning. In *IEEE Conference on Automatic Face and Gesture Recognition (FG 2018)*, pp. 409–415. Xi'an, China. DOI: 10.1109/FG.2018.00066.

N. Sadoughi and C. Busso. July 2019. Speech-driven animation with meaningful behaviors. *Speech Communication*, 110: 90–100. DOI: 10.1016/j.specom.2019.04.005.

N. Sadoughi and C. Busso. 2020. Speech-driven expressive talking lips with conditional sequential generative adversarial networks. *IEEE Transactions on Affective Computing*, To appear. DOI: 10.1109/TAFFC.2019.2916031.

N. Sadoughi, Y. Liu, and C. Busso. November 2014. Speech-driven animation constrained by appropriate discourse functions. In *International conference on multimodal interaction (ICMI 2014)*, pp. 148–155. Istanbul, Turkey. DOI: 10.1145/2663204.2663252.

N. Sadoughi, Y. Liu, and C. Busso. May 2015. MSP-AVATAR corpus: Motion capture recordings to study the role of discourse functions in the design of intelligent virtual agents. In *1st International Workshop on Understanding Human Activities through 3D Sensors (UHA3DS 2015)*, pp. 1–6. Ljubljana, Slovenia. DOI: 10.1109/FG.2015.7284885.

N. Sadoughi, Y. Liu, and C. Busso. December 2017. Meaningful head movements driven by emotional synthetic speech. *Speech Communication*, 95: 87–99. DOI: 10.1016/j.specom.2017.07.004.

K. R. Scherer. 2001. Appraisal considered as a process of multilevel sequential checking. In K. Scherer, A. Schorr, and T. Johnstone, eds., *Appraisal Processes in Emotion: Theory, Methods, Research*, pp. 92–119. Oxford University Press.

K. R. Scherer. 2005. What are emotions? and how can they be measured? *Social science information*, 44(4): 695–729.

K. R. Scherer and H. Ellgring. 2007. Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion*, 7(1): 113–130.

K. R. Scherer, E. Clark-Polner, and M. Mortillaro. 2011. In the eye of the beholder? universality and cultural specificity in the expression and perception of emotion. *International Journal of Psychology*, 46(6): 401–435.

K. Schindler, L. Van Gool, and B. De Gelder. 2008. Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural networks*, 21(9): 1238–1246.

M. Schroder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. Ter Maat, G. McKeown, S. Pammi, M. Pantic, et al. 2011. Building autonomous sensitive artificial listeners. *IEEE transactions on affective computing*, 3(2): 165–183.

S. K. Scott, N. Lavan, S. Chen, and C. McGettigan. 2014. The social life of laughter. *Trends in cognitive sciences*, 18(12): 618–620.

M. Stone, D. DeCarlo, I. Oh, C. Rodriguez, A. Stere, A. Lees, and C. Bregler. August 2004. Speaking with hands: Creating animated conversational characters from recordings of human performance. *ACM Transactions on Graphics (TOG)*, 23(3): 506–513. DOI: 10.1145/1015706.1015753.

S. Taylor, A. Kato, I. Matthews, and B. Milner. September 2016. Audio-to-visual speech conversion using deep neural networks. In *Interspeech 2016*, pp. 1482–1486. San Francisco, CA, USA. DOI: 10.21437/Interspeech.2016-483.

S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. Rodriguez, J. Hodgins, and I. Matthews. July 2017. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(4). DOI: 10.1145/3072959.3073699.

H. Tennent, S. Shen, and M. Jung. 2019. Micbot: A peripheral robotic object to shape conversational dynamics and team performance. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 133–142. IEEE.

H. D. ter Maat M., Truong K.P. 2010. How turn-taking strategies influence users' impressions of an agent. In *Intelligent Virtual Agents. IVA 2010*. Springer, Berlin, Heidelberg.

J. Tewell, J. Bird, and G. R. Buchanan. 2017. The heat is on: a temperature display for conveying affective feedback. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1756–1767.

M. Teyssier, G. Bailly, C. Pelachaud, and E. Lecolinet. 2018. Mobilimb: Augmenting mobile devices with a robotic limb. In P. Baudisch, A. Schmidt, and A. Wilson, eds., *The 31st Annual ACM Symposium on User Interface Software and Technology, UIST 2018, Berlin, Germany, October 14-17, 2018*, pp. 53–63. ACM. https://doi.org/10.1145/3242587.3242626. DOI: 10.1145/3242587.3242626.

M. Teyssier, G. Bailly, C. Pelachaud, and E. Lecolinet. 2020. Conveying emotions through device-initiated touch. *IEEE Trans. Affective Computing*, 01.

N. M. Thalmann, P. Kalra, and M. Escher. 1998. Face to virtual face. *Proceedings of the IEEE*, 86(5): 870–883.

K. Thórisson. 1997. Layered modular action control for communicative humanoids. In *Computer Animation'97*. IEEE Computer Society Press, Geneva, Switzerland.

O. Torres, J. Cassell, and S. Prevost. 1997. Modeling gaze behavior as a function of discourse structure. In *In Proceedings of the First International Workshop on Human-Computer Conversations*.

N. Tsapatsoulis, A. Raouzaiou, S. Kollias, R. Cowie, and E. Douglas-Cowie. 2002. Emotion recognition and synthesis based on MPEG-4 FAPs in MPEG-4 facial animation. In I. S. Pandzic and R. Forcheimer, eds., *MPEG4 Facial Animation - The standard, implementations and applications*. John Wiley & Sons.

L. Valbonesi, R. Ansari, D. McNeill, F. Quek, S. Duncan, K. McCullough, and R. Bryll. September 2002. Multimodal signal analysis of prosody and hand motion: Temporal correlation of speech and gestures. In *European Signal Processing Conference (EUSIPCO 02)*, pp. 75–78. Tolouse, France.

A. Vidal, A. Salman, W.-C. Lin, and C. Busso. October 2020. MSP-face corpus: A natural audiovisual emotional database. In *ACM International Conference on Multimodal Interaction (ICMI 2020)*. Utrecht, The Netherlands.

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.

K. Wada, T. Shibata, K. Sakamoto, and K. Tanie. 2006. Long-term interaction between seal robots and elderly people — robot assisted activity at a health service facility for the aged —. In K. Murase, K. Sekiyama, T. Naniwa, N. Kubota, and J. Sitte, eds., *Proceedings of the 3rd International Symposium on Autonomous Minirobots for Research and Edutainment (AMiRE 2005)*, pp. 325–330. Springer Berlin Heidelberg, Berlin, Heidelberg.

H. Wallbott. 1998. Bodily expression of emotion. *European Journal of Social Psychology*, 28: 879–896.

C. Willemse, A. Toet, and J. van Erp. 2017. Affective and behavioral responses to robot-initiated social touch: toward understanding the opportunities and limitations of physical contact in human–robot interaction. *Frontiers in ICT*.

L. Williams. August 1990. Performance-driven facial animation. *Computer Graphics*, 24(4): 235–242. DOI: 10.1145/1185657.1185856.

S. Yohanan and K. MacLean. 2012. The role of affective touch in human-robot interaction: Human intent and expectations in touching the haptic creature. *Int J of Soc Robotics*, 4: 163–180.

Y. Yoon, W.-R. 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. IEEE.

F. Yunus, C. Clavel, and C. Pelachaud. July 2019. Gesture class prediction by recurrent neural network and attention mechanism. In *International Conference on Intelligent Virtual Agents (IVA 2019)*, pp. 233–235. Paris, France. DOI: 10.1145/3308532.3329458.

A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6): 82–88.

A. Zadeh, P. Liang, J. Vanbriesen, S. Poria, E. Tong, E. Cambria, M. Chen, and L.-P. Morency. July 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *ACM Association for Computational Linguistics (ACL 2004)*, volume 1, pp. 2236–2246. Melbourne, Australia. DOI: 10.18653/v1/P18-1208.