INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2021.1898851

Examining the Use of Nonverbal Communication in Virtual Agents

Isaac Wang and Jaime Ruiz

Department of Computer and Information Science and Engineering, University of Florida, Gainesville, Florida, USA

ABSTRACT

Virtual agents are systems that add a social dimension to computing, often featuring not only natural language input but also an embodiment or avatar. This allows them to take on a more social role and leverage the use of nonverbal communication (NVC). In humans, NVC is used for many purposes, including communicating intent, directing attention, and conveying emotion. As a result, researchers have developed agents that emulate these behaviors. However, challenges pervade the design and development of NVC in agents. Some articles reveal inconsistencies in the benefits of agent NVC; others show signs of difficulties in the process of analyzing and implementing behaviors. Thus, it is unclear what the specific outcomes and effects of incorporating NVC in agents and what outstanding challenges underlie development. This survey seeks to review the uses, outcomes, and development of NVC in virtual agents to identify challenges and themes to improve and motivate the design of future virtual agents.

1. Introduction

Virtual agents are computer systems that strive to engage with users on a social level, through the use of technologies such as natural language interfaces and digital avatars. While traditional dialogue systems enable natural user interaction by allowing users to speak and converse with a computer through natural discourse (Jurafsky & Martin, 2000), virtual agents have the ability to go a step further by supplementing the system with a visual representation, or avatar. For example, embodied conversational agents extend the capabilities of dialogue systems through an embodiment (Cassell, 2001), which helps an agent take on a social role in interaction as users treat the agent more as another person or individual (Reeves & Nass, 1996; Sproull et al., 1996; Walker et al., 1994).

However, people communicate through more than just speech. For instance, we use gesture to illustrate ideas and provide information (Kendon, 2004; McNeill, 1992), gaze to show responsiveness and direct attention (Frischen et al., 2007), and facial expressions to convey emotion and mood (Ekman, 1993). While dialogue systems are unable to fully take advantage of these modalities, virtual agents can emulate these behaviors through an embodiment and thus leverage the full multimodality of human communication (Cassell, 2001).

The use of nonverbal communication (NVC) in virtual agents is practically as old as agents themselves (Cassell, 2001). Although rudimentary compared to more recent examples, early virtual agents already exhibited different nonverbal behaviors in addition to speech. For instance, the Rea agent (Cassell et al., 1999) was able to communicate through speech and gesture, pointing at objects to refer to them and making sweeping motions to pass the speaking floor back and forth. Other agents focused on other forms of NVC such as facial expressions to express emotion and adorn their interaction with affective content (Becker et al., 2004; Cassell & Thorisson, 1999). With advances in technology, more recent agents feature complex models of NVC (Andrist et al., 2012b; Cafaro et al., 2016; Pelachaud, 2017) and use a combination of different types of behaviors (DeVault et al., 2014; Gratch, Wang, Gerten et al., 2007; Traum et al., 2008).

Despite the number of agents that incorporate some form of NVC, to our knowledge, there are few literature reviews that focus on how NVC is used in agents, and none that explore the roles and outcomes of different types of behaviors and the efforts required to develop an agent that uses NVC. Previous surveys (Allbeck & Badler, 2001; André & Pelachaud, 2010; Nijholt, 2004) have presented an overview of the different behaviors that agents have employed, but do not elaborate on the inherent difficulties nor identify the outcomes of using such behaviors. Other articles focus on how the appearance of an agent, which can be considered a type of NVC (Argyle, 1988), affects how users perceive it (Baylor, 2009, 2011). However, these articles do not focus on more explicit nonverbal *behaviors*, which are what we seek to study in our review.

More relevant are the surveys of pedagogical agents, which detail the effectiveness of agents in teaching scenarios (Clarebout et al., 2002; Heidig & Clarebout, 2011; Johnson et al., 2000; Krämer & Bente, 2010), but these surveys do not focus on the use of NVC, or only mention it in passing. However, a few papers hint at the existence of issues in virtual agent NVC, such as the inconsistency of effectiveness in teaching scenarios (Baylor & Kim, 2008; Frechette & Moreno, 2010) and the time-consuming nature of defining nonverbal behaviors (Rehm & André, 2008).

CONTACT Isaac Wang wangi@ufl.edu Department of CISE, University of Florida, 432 Newell Dr., Gainesville, FL 32611, USA

Thus, the goal of this literature review is to explore the use of NVC in virtual agents and identify the challenges with incorporating NVC in virtual agents, from both the interaction and development standpoints. While agents can feature a range of embodiments from none (e.g., dialogue systems) to physical, real-world embodiments (e.g., robots), for our review, we focus specifically on virtual agents that feature a *digital* avatar, or visual representation of that agent.

We begin our survey in Section 3 by highlighting the role NVC takes in human-human interactions and then continue in Section 4 by comparing how NVC is likewise emulated in virtual agents. In Section 5, we detail the effects and outcomes NVC has on human-agent interactions, and then, in Section 6, we describe how agent systems that incorporate NVC are designed and developed. We follow this in Section 7 with a discussion of the primary challenges and themes identified in the literature. We then conclude in Section 8 by summarizing our findings, with the intent of motivating continued research into the creation of virtual agents that use NVC.

2. Methodology

To conduct the paper search portion of this literature survey, we utilized the methodology by Kitchenham et al. (2009), who present a set of guidelines for conducting a systematic literature review. For our review, we focused on adapting their strategy for planning research questions and identifying relevant papers. In their process, they recommend starting with establishing a set of research questions/goals to motivate the initial search. To start, we came up with a number of different research questions based on the topic of incorporating NVC in virtual agents. These are listed below:

(1) How is NVC used in virtual agents?
   (a) Specifically, what types of NVC have been implemented?
   (b) How do the outcomes and functions of NVC in agents compare with natural human interaction?

(2) What challenges are there to creating virtual agents that use NVC?
   (a) How is NVC implemented in virtual agent systems?
   (b) How do researchers understand and define nonverbal behaviors?

(3) How is mirroring/mimicry utilized in virtual agents that display NVC?

Based on these questions, we derived a set of keywords and search combinations that would yield relevant results. For instance, to identify implementations of NVC in agents, we searched for "nonverbal behavior" and "virtual agent," which yielded a list of articles relating to virtual agents that used some form of NVC. Key search terms included: *Nonverbal communication, gesture, facial expression, multimodal, virtual agent, mirroring, mimicry, chameleon effect*.

These keyword combinations and synonyms (e.g., "virtual agent," "embodied conversational agent," and "virtual human") were inputted into Google Scholar to generate a broad listing of articles aggregated from multiple sources. For each search result, we looked through at least the first 10 result pages (up to 15, based on if we encountered many repeated entries in the first 10 pages) for each of the search terms and identified papers relevant to our research questions. This yielded 79 papers from the initial search protocol.

After this search, we also further reviewed the related work and other citations from each selected paper. From these, the relevant articles were added to our paper list. Likewise, these newly added papers were also reviewed to find additional relevant papers, what Kitchenham et al. (2009) refer to as "snowballing." Using this method, 43 papers were added to the list.

Additionally, papers from relevant authors as identified during the search process (and noted as commonly cited, during the snowballing method) were also reviewed for inclusion in our literature review. We searched for these authors on Google Scholar and their relevant works were also evaluated for inclusion. 22 papers from related authors (that were not already included in the list) were added.

During the search process, papers suggested by collaborators were also added to the list. The resulting number of papers identified through each method is listed in Table 1, for a total of 150 papers selected to be in this literature review.

3. NVC in human-human interactions

To understand how to effectively implement nonverbal behaviors in virtual agents, we first examine how they are used in humans. Nonverbal cues are used as an additional channel of communication that can be used alongside speech to modify what is being said, adding meaning and richness (Argyle, 1988; Kendon, 2004; Quek et al., 2002). In this section, we briefly describe the relevant main benefits of NVC in human interactions as the key motivation and rationale behind the use of NVC in virtual agents. Although there are many forms of nonverbal behavior, we discuss the primary ones that are most relevant to communication and common among the virtual agents identified through our survey.

3.1. Gesture

One of the most evident forms of NVC is gesture. Gestures are motions and poses primarily made with the arms and hands (or sometimes other body parts) in order to communicate, often while speaking (Figure 1). Gesture is used by humans in everyday conversation to refer to objects and add expressiveness to language by demonstrating events and actions (Kendon, 2004). Kendon (1988) and McNeill (1992) further separate gestures into five different categories with differing communicatory functions. They range from simple *beat* gestures (rhythmic motions that go along with words) to *emblem* gestures (that have their own meaning and can fully replace words). These can add dynamics to a conversation, driving the point through a *beat* gesture or signifying approval through an *emblematic* thumbs-up. These gestures co-exist

Table 1. Paper identification methods.

Identification Method	# of Papers Identified
Search Protocol	79
Snowballing	43
Relevant Authors	23
Suggested by Collaborators	10
Total	**155**

with speech, adding complementary or redundant information. By annotating and quantitatively analyzing videos of people communicating, Quek et al. (2002) show how gesture and speech form a temporal and co-expressive relationship with one another. Gestures occur at key points in time while speaking, such as during speech or specifically during a verbal pause.

While these examples highlight the descriptive use of gesture in the contexts of narration and conversation, gestures also play a more functional role in delivering information. Due to their spatial nature, gestures have the ability to describe spatial features, allowing a person to illustrate imaginary objects and spaces (Alibali, 2005; Bergmann, 2006; Kendon, 2004). They also facilitate conversational grounding, i.e. the process of establishing mutual understanding when communicating (Clark & Brennan, 1991; Clark & Wilkes-Gibbs, 1986; Dillenbourg & Traum, 2006). Gestures also help manage the flow of conversation by functioning as a signal for turn-taking and also providing a "back-channel" for communicating attention (Duncan & Niederehe, 1974; Young & Lee, 2004), such as a nodding head gesture. The use of gesture in teaching is also associated with comprehension and information recall (Cook & Goldin-Meadow, 2006; Goldin-Meadow & Alibali, 2013; Novack & Goldin-Meadow, 2015).

The HCI field has also examined the use of gesture to communicate, with the express intent to build gestural interaction systems. For instance, in a study looking at how people use gestures to describe objects and actions, Grandhi et al. (2011) saw how people used pantomime to directly paint an imaginary image of the action (i.e., holding imaginary objects and pretending to perform a task) rather than attempt to use abstract gestures to describe the task or object. Other research focusing on shared visuals and workspaces between people also highlight the usefulness of non-verbal communication to support coordination between humans (Fussell et al., 2000, 2004; Gergle et al., 2004; Kraut et al., 2003). In these cases, *deictic* gestures (i.e., pointing) are used to direct shared attention between people and signal toward specific objects or areas when speaking.

3.2. Gaze

As a nonverbal signal, gaze has both interpersonal effects as well as practical functions. For example, mutual gaze or eye contact can signal likability or intimacy (Argyle & Dean, 1965; Cook, 1977), but gazing for too long can increase discomfort or awkwardness (Cook, 1977). Mutual gaze (eye contact between individuals) has also been shown to have an impact on a person's credibility (Beebe, 1976). On a more functional level, shared gaze, or joint attention, has been shown to increase performance in spatial tasks between collaborators (Brennan et al., 2008). This is due to the ability for gaze to both convey attention and direct attention (Frischen et al., 2007), with similar effect to deictic gestures.

There are also conversational functions associated with gaze behavior. While speaking, a person may use gaze to nonverbally manage the conversation (Duncan, 1972). The use of key gaze signals (such as averting the eyes and then returning to a mutual gaze) can be used when turn-taking to pass the speaking floor to someone else or maintain control of the floor. When listening, gaze can be indicative of a person's level of interest and attention (Bavelas & Chovil, 2006; Bavelas et al., 2002). A speaker would often gaze away while speaking but engage in brief moments of mutual gaze in order to

Figure 1. A man naturally gesturing with his hands while speaking to emphasize the importance of a concept. (*Public domain image*).

"checkup" on the listener. The listener would react to the gaze by giving an affirmative response (whether verbal or gestural), to which the speaker would resume. Likewise, gaze can also be a measure of listener comprehension (Beebe, 1976).

3.3. Facial expression

Facial expressions are often used by humans to convey emotion or affect. While facial expressions are not the only channel for communicating emotion, they are among the strongest indicators (Ekman, 1993). Research by Ekman (1992) has shown how, as humans, we have at minimum a basic set of emotions (ranging from anger to surprise) with virtually universal facial expressions to convey them (Ekman et al., 1987). In contrast, Horstmann (2003) argues that expressions signal more than just emotion, but also reveal the person's intent and can even communicate desired actions. Facial expressions are also readily interpretable by people (H. L. Wagner et al., 1986), as well as by facial recognition systems (Bartlett et al., 1999). Thus, facial expressions provide a rich source of information that is both readily available and recognizable.

3.4. Proxemics

There are other behaviors which also communicate but are more focused around the whole body and its positioning. One such focus is on proxemics, or the study of the interpersonal distance between individuals and how that form of body language is interpreted. Hall (1966) laid out four progressive levels of interpersonal distance, ranging from a close *intimate* distance to a far *public* distance. People naturally interact at these different distances based on their interpersonal relationship (Willis, 1966), for instance, talking with friends would be at a personal distance (around 2–5 feet) but likely not at an intimate distance (less than 2 feet). When meeting someone for the first time, how close they come can also express different personality traits, such as friendliness or extroversion (Patterson & Sechrest, 1970).

3.5. Posture

Postures have been shown to express a number of different attitudes and meanings. A posture with arms folded or legs crossed can signify being "closed-off" or less inviting to social interaction (Argyle, 1988). Mehrabian (1969) showed how postures directly related to feelings of assertion and dominance, based on the perceived relaxation of a pose. They also showed how likability and attentiveness can be conveyed through body language such as leaning toward another person, sometimes in conjunction with eye contact and closer proximity (Mehrabian, 1972).

Similar to facial expression, posture also has the ability to communicate emotional state. A study by Dael et al. (2012) showed how specific patterns of body movements and poses commonly expressed specific emotions and that people were readily able to differentiate between them. The level of intensity for an emotion can also be conveyed through the body. Wallbott (1998) studied how the amplitude/level of energy in motions and body poses can reflect the intensity of the emotion portrayed. Posture can also be a measure of rapport (Kendon, 1970), more recently studied through the psychological effect of mirroring, where a person mimics the body poses, gestures, or even general attentiveness of another, often subconsciously (Lafrance & Broadbent, 1976). In doing so, this is interpreted as a sign of likability or willingness to cooperate (Chartrand & Bargh, 1999).

3.6. Behavioral mirroring

One particular nonverbal behavior we want to introduce here is mirroring. Mirroring is different from other behaviors in that it is not technically a "type" of NVC. Instead, mirroring is when one person subconsciously mimics the behavior of another (Kendon, 1970). The act may be verbal or nonverbal; however, for this paper, we refer to behavioral mirroring in the nonverbal sense. Mirroring has many benefits in human-human communication, including increasing likability (Chartrand & Bargh, 1999; Duy & Chartrand, 2015; Jacob, Gueguen, Martin, & Boulbry, 2011), rapport (Chartrand & Bargh, 1999; Duy & Chartrand, 2015; Kendon, 1970; Lafrance & Broadbent, 1976; Lakin & Chartrand, 2003), and persuasion (Jacob et al., 2011; van Swol, 2003).

Kendon (1970) was one of the earliest to study mirroring, calling it "interactional synchrony" that naturally occurred between individuals. Lafrance and Broadbent (1976) noted that posture mimicry (e.g., leaning forward after a speaker leans forward) may be a subconscious signal to convey a listener's attention. This signal, when (subconsciously) received back by the speaker, can increase feelings of rapport. LaFrance (1985) even refers to mirroring as "*an obvious yet unobtrusive indicator of openness to interpersonal involvement*."

These two studies were observational; Chartrand and Bargh (1999) were the first to investigate mirroring in an experimental setting. In their study, participants described photos to another individual (a researcher who functioned as a confederate). The confederate would either mimic the participant's body language (e.g., gesture and posture shifts) or not. Participants who interacted with the mimicking confederate rated their interaction as smoother, with a similar increase in likability and empathy.

Mirroring can even influence behavior. In a study by Jacob et al. (2011), customers who interacted with a retail clerk that mimicked their behaviors rated the clerk as more likable and having more influence. The customers who interacted with the mirroring clerk were also more likely to spend more. Van Baaren et al. (2004) even showed how mirroring may even have positive social effects. They conducted a study where participants interacted with either a mirroring or non-mirroring confederate. The confederate would "accidentally" drop a few pens on the ground. Participants in the mirroring condition were more likely to help the confederate pick up the pens. Cook and Goldin-Meadow (2006) studied how mirroring can impact learning in children. Children who were taught by an instructor who used gesture were more likely to use the same gestures when explaining the concepts they learned. The children who mirrored the gesture also scored

higher on a test versus children who did not mirror, despite both receiving the same instruction.

3.7. Complex NVC

Furthermore, an important point about these nonverbal behaviors is that they are not all independent – different behaviors are often combined to form more complex nonverbal expressions. A person might, for example, both gaze at another while gesturing toward them to signal that they are passing the speaking floor (Duncan, 1972; Duncan & Niederehe, 1974). In addition, the act of communicating understanding involves gaze, facial expression, and even posture (Clark & Brennan, 1991; Duncan & Niederehe, 1974; Young & Lee, 2004). Proximity can also be used as an indicator of willingness to interact (Argyle, 1988), such as walking toward someone but in conjunction with gaze and facial expression to communicate their intention. In contrast, two types of NVC may even interact, such as with mutual gaze and proximity (Argyle & Dean, 1965). Gaze may improve levels of intimacy between two people but can be affected by their proximity (thus requiring a balance between the two in order to create the desired effect).

4. NVC in virtual agent interactions

In the previous section, we highlighted different types of NVC used in human-human interactions. These encompassed a wide range of nonverbal cues, from gesture to proxemics. Likewise, virtual agents have implemented similar cues in order to emulate natural human behavior. These agents have been created for a number of different application areas, including teaching (Andrist, 2013; Baylor & Kim, 2008, 2009; Noma et al., 2000; Rickel & Johnson, 1999, 2000), coaching (Anderson et al., 2013; Bergmann & Macedonia, 2013; Kang et al., 2008), healthcare (DeVault et al., 2014; Hirsh et al., 2009; Kang et al., 2012), military settings (Kenny et al., 2007; Lee et al., 2007; Traum et al., 2008), as general conversation partners (Buschmeier & Kopp, 2018; Gratch, Wang, Okhmatovskaia et al., 2007; Hartmann et al., 2006; Pelachaud, 2005a), and as virtual assistants (Cassell et al., 1999; Matsuyama et al., 2016; Theune, 2001). In our literature search, we identified papers describing agents that implement each of the NVC types detailed in the previous section. In this section, we give examples of how virtual agents emulate each type of behavior. We start by summarizing the types of NVC used in agents and how they compare against human-human interactions, before later describing the overall goals and effects of agent NVC in Section 5.

4.1. Gesture

Gesture is one of the most predominant forms of NVC used in agents. Recall that, in humans, gesture has many purposes ranging from illustrating words and ideas to directing attention and referring to objects. We see these same purposes featured in virtual agents. Like with human-human interactions, gesture is primarily used to accompany speech. The Rea agent, created by Cassell et al. (1999) to function as a virtual real estate agent, is one of the earliest to use gestures for this purpose. As she interacts with users, she will nod (a head gesture) to signal attention and use beat gestures to emphasize specific words when she speaks.

Similarly, the Greta agent (Niewiadomski et al., 2009; Pelachaud, 2017; Poggi et al., 2005) in Figure 2 is a multipurpose conversational agent that is designed for multipurpose applications (ranging from interviews to coaching). This allows her to be adapted and used for different research goals. One of the key features of the Greta agent is that she performs different gestures when speaking to the user. Like with human-human communication, Greta's use of gestures is for added expressivity, but also for the goal of increasing the agent's believability in interactions. The Max agent (Becker et al., 2004) also performs gestures while speaking to increase believability and realism.

On a different note, the Steve agent (Rickel & Johnson, 1999) also uses gestures when speaking, but instead in teaching and training situations. Unlike the previous agents that use gesture to primarily make conversations more expressive, Steve takes a more functional approach. Steve uses deictic gestures while referring to specific objects in his environment as well as demonstrative gestures and direct actions to directly show the user how to perform a task. Traum et al. (2008) also had their agent gesture toward objects when referring to them. These examples highlight how agents, much like humans, are able to use gesture to help in completing tasks.

4.2. Gaze

Gaze is also commonly emulated in virtual agents. In particular, agents focus on implementing both mutual gaze and deictic gaze to convey their attention and refer to objects, respectively. Andrist, Mutlu et al. (2012a, 2013) describe the development of agents that employ mutual gaze and gaze aversions in a teaching setting to control the conversation and help increase feelings of affiliation in users. Likewise, an agent by Lee et al. (2007) uses a model of mutual gaze in

Figure 2. The Greta agent communicating through gesture. Image from https://github.com/isir/greta. Reproduced with permission.

conversation to show when the agent is paying attention and listening intently. More recently, the Ellie agent by DeVault et al. (2014) and the SARA agent by Matsuyama et al. (2016) use gaze in combination with other types of NVC to build rapport. Lance and Marsella (2009, 2010), Lance & Marsella (2008) go even further by using gaze aversion to convey emotion in agents.

Andrist et al. (2017) also created an agent that used gaze to coordinate and direct attention. A user was asked to collaborate with the agent to build different types of sandwiches. The agent would use gaze in conjunction with speech to refer to ingredients. In that same vein, Pejsa et al. (2015) describe the development of an agent that uses both mutual and deictic gaze to coordinate movement and show interest in both the user as well as objects in the environment. Overall, as with gesture, gaze is well-utilized in agents with much of the same purposes as in human-human interactions.

4.3. Facial expression

A number of agents have the ability to display different facial expressions. Just like with human-human interactions, facial expressions in virtual agents are used to convey emotion, or affect. As an early example, Gandalf (Cassell & Thorisson, 1999) makes different faces while answering users' questions, such as smiling while making a joke. A more advanced agent, Max, by Becker et al. (2004) has an internal model of emotion that enables him to smile when happy or even look angry when annoyed. Greta (Poggi et al., 2005), as mentioned earlier, is also expression-capable and can even use them in conjunction with other nonverbal behaviors.

These facial expressions help give an agent its own kind of personality. Cafaro et al. (2016) describe their Tinker agent, designed to function as a guide for museum exhibits. They focus on making sure that the user's first impression of the agent is a positive and friendly one, so that they continue to interact with the agent. Expression is one of the behaviors that Tinker uses to convey such a personality. McRorie et al. (2012) created four different agents featuring different expressions, enabling each of them to express their own personality.

On the healthcare side, the virtual therapist by Grolleman et al. (2006) focuses on facial expressions to try and increase feelings of empathy and acceptability. The Ellie agent (DeVault et al., 2014) mentioned earlier also uses expression alongside gaze with the goal of increasing rapport. Kang et al. (2012) focused on increasing intimacy with an agent, which is important for users to feel comfortable when disclosing sensitive information to a virtual counselor. As a whole, the use of expression enables agents to engage users on an affective level, making them more capable social actors.

4.4. Proxemics

Although not as common, a few agents exist that follow the rules of proxemics when interacting with others. The museum guide Tinker (Cafaro et al., 2016), mentioned earlier, uses more than just facial expression to engage users. Tinker also obeys the rules of interpersonal distance, ensuring that the interaction does not feel awkward by invading in intimate space or staying too far away to communicate. Edith (Andrist, Leite et al., 2013), an agent designed to interact with groups of children, typically stays within an acceptable social distance when speaking to the whole group but moves closer as needed to engage with an individual. The chat agents by Isbister et al. (2000) also employs similar group dynamics. When engaging multiple people, the agents will move back to ensure that they are seen by all parties. In contrast, the agents will turn and move closer when their speech is directed to an individual. Similarly, Pejsa et al. (2017) developed a footing model for agents to orient and engage users in virtual environments. Responding to proxemic social cues is also a way for agents to come across as more personal and aware (Garau et al., 2005).

In humans, the rules of proxemics also differ based on cultural norms. Distinct cultures have varying comfort levels regarding interpersonal distance. Virtual agents have also been designed that mimic these behaviors. Kistler et al. (2012) created agents that exhibited proxemic behaviors for both individualistic and collectivistic cultures. As a result, users had different responses to the agents based on their own cultural standards.

4.5. Posture

As with proxemics, posture is also not as commonly used compared to other NVC types such as gaze and gesture. However, there are some full-body agents that do assume different postures while interacting with a user. The therapist agent Ellie (DeVault et al., 2014) is one agent that increases rapport through appropriate body postures (Figure 1), in an attempt to emulate what is done by real therapists. Gratch et al. (2006), Huang et al. (2011), and Kang et al. (2008) also focus on rapport in conversation and interview settings. Their agents use posture shifts alongside other types of NVC to increase users' feelings of rapport and make the agent feel more natural.

4.6. Behavioral mirroring

Mirroring is also not as prominent in virtual agents. However, researchers have developed a few agents that mirror the nonverbal behavior of users. For example, the Rapport Agent by Gratch et al. (2006) implements posture, gaze, and head motion (nod and shake) mimicry as a form of listening feedback. The goal is that the inclusion of these behaviors increases a user's sense of rapport with the agent, much like in humans. Similarly, Stevens et al. (2016) created an agent that would mimic a user's facial expressions while speaking, to increase feelings of lifelikeness and likability. Other agents emulate the expressiveness of gestures (Bevacqua et al., 2006; Caridakis et al., 2007), mimic head motions (Bailenson & Yee, 2005; Bailenson et al., 2008), and gestures (Castellano et al., 2012).

5. Effects of NVC in virtual agents

In the previous sections, we highlighted the benefits of NVC in humans and how agents attempt to emulate this behavior. In this section, we focus on how the NVC produced by agents affects users, specifically, the different roles it plays and the corresponding outcomes and effects. For example, many agents use facial expressions with the goal of conveying emotion, but do users actually pick up on these signals and correctly identify the agent's emotion? This section aims to answer this question and understand the outcomes of NVC in virtual agents. Table 2 provides a summary of the different purposes and objectives of agent NVC identified in our review. In the following subsections, we describe these in detail, highlighting the outcomes and effects that NVC has on users.

5.1. Managing conversation

To start, as in humans, NVC in agents can be used to manage a conversation, which affects how users perceive an agent's engagement and helpfulness. For instance, Andrist, Leite et al. (2013) showed how NVC can be effective for managing conversation in groups. They conducted a study where an agent would try to manage a group of children. The agent would use multimodal cues (including gaze, gesture, and proxemics) to pass the speaking floor around so that each child would have an equal opportunity to contribute. The study showed that an agent that includes all three cues resulted in a more evenly managed group with lower variance in turns taken. Compared to an agent that only used vocal cues or a subset of the nonverbal cues, the agent that used all three was most successful in managing the conversation without decreasing the enjoyment of the agent.

Research has also shown how NVC in agents can also be effective in communicating feedback such as attention. For example, Bailenson et al. (2002) showed that the ability for head movements to communicate gaze can be transferred into virtual agents with similar effects. They represented users with motion-tracked avatars in a virtual environment and showed that the use of communicative head movements still enriched interaction and decreased the proportion of speech needed (as gaze could be used to nonverbally communicate attention and intention). Turning the problem around, Buschmeier and Kopp (2018) created an "attentive" agent that responded to the **user's** levels of attention. Their agent that was able to adapt its communication based on human attentiveness and verbal/nonverbal feedback (e.g., head gestures). They compared this agent against one that did not respond to these inputs. Surprisingly, they found that humans naturally produced more of these feedback behaviors for the attentive agent. This points to the possibility that, at some level, users were aware of the fact that the agent could understand their behaviors, and thus communicated more in these modalities. In their study, subjective ratings reflected that users were aware of the agent's capabilities: they rated the agent higher in terms of understanding them and feeling attentive. These are key examples of how agents can effectively use NVC to perform different conversational functions. However, recall that agents also perform nonverbal cues with less functional, and more expressive purposes.

An early study by Cassell and Thorisson (1999) aimed to understand which type of cues were more important in human-agent interactions. They separated nonverbal feedback into two categories: *emotional*, which focuses more on conveying affect (through facial expressions and other body language), and *envelope*, which focuses on conversational functions (through gaze, gestures, etc.). By their definition, envelope feedback is feedback that does not add to a conversation, it merely coordinates it. As an example, averting one's gaze to signal that they are taking a turn, or expressing attention while the other is speaking would be considered forms of envelope feedback. They conducted a study asking participants to interact with an agent that exhibited content-only feedback (speech without NVC), content with emotional feedback, and content with envelope feedback. Users rated the agent with envelope feedback as more helpful and efficient than the one with emotional feedback.

5.2. Expressing a unique personality

In contrast to this work is the idea that emotional feedback/behaviors are also important. Agents often need to focus on changing users' perceptions of them. For instance, personality is important in virtual agents, especially when agents are targeted toward healthcare and therapy applications (DeVault et al., 2014). An agent should ideally have an appropriate personality and behavior for its given context. Giving a computer system a face allows it to take on a social role (Reeves & Nass, 1996). In doing so, people naturally view an agent in light of different social attributes: for instance, people view a computer with a face as more likable (Walker et al., 1994) and often ascribe specific personalities to it (Sproull et al., 1996). They change their attitude toward the more human-like interface than a bare system. By extension, how an agent *behaves*, both verbally and nonverbally, can have a large influence on its perceived personality (Cafaro et al., 2012).

A number of studies have focused on studying the interplay between NVC and personality. Although personality and emotional traits are primarily delivered through speech and facial expression, in some cases, even behavior alone is enough to convey a sense of emotion and personality (Clavel et al., 2009). As an example, Neff et al. (2010) looked at correlations between different behavioral characteristics against perceptions of introversion and extroversion. They found that tweaking different gesture parameters, such as speed and the range of motion used, changed the perceived personality of an agent. In their study, agents that displayed gestures where limbs stayed close to the body were rated as more introverted compared to gestures that extended further. Performing gestures more often also increased senses of extroversion. McRorie et al. (2012) extended this to show how additional characteristics of face and gesture (such as gesture speed, spatial volume, energy, etc.) can also express the traits of extraversion, psychoticism, and neuroticism. A user's perceptions of extraversion and affiliation are also affected by the inclusion of smile, gaze, and proxemics (Cafaro et al., 2012). When studying first impressions of

Table 2. Papers grouped by type and purpose of virtual agent NVC.

		Type of Nonverbal Communication					
		Gesture	Gaze	Facial Expression	Proxemics	Posture	Mirroring
Purpose or Objective	Conversational Function	Andrist (2013), Cassell et al. (1999), Matsuyama et al. (2016), Nguyen and Masthoff (2009)	Andrist (2013), Andrist, Mutlu et al. (2013), Cassell et al. (1999), Isbister et al. (2000), Lee et al. (2007), Matsuyama et al. (2016), Poggi et al. (2000)	Cassell et al. (1999), Matsuyama et al. (2016)	Andrist, Leite et al. (2013), Isbister et al. (2000), Pejsa et al. (2017)		
	Attention/ Feedback	Buschmeier and Kopp (2018), Cassell and Thorisson (1999), Gratch et al. (2006), Gratch, Wang, Gerten et al. (2007)	Buschmeier and Kopp (2018), Cassell and Thorisson (1999), Garau et al. (2005)	Buschmeier and Kopp (2018)	Garau et al. (2005)		Bailenson and Yee (2005), Gratch, Wang, Gerten et al. (2007), Huang et al. (2011), Stevens et al. (2016)
	Personality	Anderson et al. (2013), Becker et al. (2004), Hartmann et al. (2005), Hartmann et al. (2006), McRorie et al. (2012), Neff et al. (2010). Poggi et al. (2005). Theune (2001)	Cafaro et al. (2016), Cafaro et al. (2012). McRorie et al. (2012), Poggi et al. (2005)	Anderson et al. (2013), Becker et al. (2004), Cafaro et al. (2016), Cafaro et al. (2012), McRorie et al. (2012), Poggi et al. (2005), Theune (2001)	Cafaro et al. (2016), Cafaro et al. (2012), Poggi et al. (2005)	Neff et al. (2010)	
	Copresence		Bailenson et al. (2005), (2006), Garau et al. (2005), Guadagno et al. (2007)		Garau et al. (2005), Huang et al. (2011)	Huang et al. (2011)	Bailenson and Yee (2005)
	Rapport	DeVault et al. (2014), Gratch et al. (2006), Gratch, Wang, Gerten et al. (2007), Huang et al. (2011), Matsuyama et al. (2016)	DeVault et al. (2014), Gratch et al. (2006), Gratch, Wang, Gerten et al. (2007), Huang et al. (2011), Kang et al. (2008). Matsuyama et	DeVault et al. (2014), Huang et al. (2011), Matsuyama et al. (2016)		DeVault et al. (2014), Gratch et al. (2006), Gratch, Wang, Gerten et al. (2007), Huang et al. (2011), Kang et al. (2008)	Chartrand and Bargh (1999), Gratch et al. (2006), Gratch, Wang, Gerten et al. (2007), Huang et al. (2011)
	Trust/Intimacy	Buisine et al. (2004), Kang et al. (2012), Traum et al. (2008)	Andrist et al. (2012a), Kang et al. (2012), Traum et al. (2008)	Kang et al. (2012)	Kistler et al. (2012)	Traum et al. (2008)	Bailenson and Yee (2005), Bailenson et al. (2008), Verberne et al. (2013)
	Emotion/ Empathy	Anderson et al. (2013), Becker et al. (2004), Hartmann et al. (2005), (2006), Nguyen and Masthoff (2009), Pelachaud (2005b), Poggi et al. (2005)	Nguyen and Masthoff (2009), Pelachaud (2005b), Poggi et al. (2005)	Anderson et al. (2013), Becker et al. (2004), Becker et al. (2007), Cassell and Thorisson (1999), Nguyen and Masthoff (2009), Niewiadomski et al. (2008), Pelachaud (2005b, 2009a), Poggi et al. (2005)			Castellano et al. (2012), Krämer et al. (2013)
	Collaboration/ Coordination	Andrist, Leite et al. (2013), Kopp and Wachsmuth (2004), Krämer et al. (2003), Rickel and Johnson (2000)	Andrist et al. (2017), Andrist, Leite et al. (2013), Andrist et al. (2012a), Bailenson et al. (2006), Rickel and Johnson (2000)	Rickel and Johnson (2000)	Andrist, Leite et al. (2013)	Rickel and Johnson (2000)	
	Learning/ Training	Anderson et al. (2013), Baylor and Kim (2008, 2009), Bergmann and Macedonia (2013), Buisine et al. (2004), Buisine and Martin (2007), Frechette and Moreno (2010), Kenny et al. (2007), Noma et al. (2000), Rickel and Johnson (1999)	Andrist et al. (2012a), Buisine et al. (2004), Buisine and Martin (2007), Kenny et al. (2007), Noma et al. (2000), Rickel and Johnson (1999)	Anderson et al. (2013), Baylor and Kim (2008, 2009), Buisine and Martin (2007), Kenny et al. (2007), Noma et al. (2000)		Kenny et al. (2007), Rickel and Johnson (1999)	

virtual agents, Cafaro et al. (2012) found that while proximity mainly affected perceptions of extraversion, smiling behavior dominated positive impressions of friendliness, and gaze had a smaller influence on personality.

Personality also affects the perceived warmth and competence of an agent. When analyzing behaviors to implement in agents, Biancardi et al. (2017a) analyzed a corpus of human-human interactions and found a correlation between gesture use and personality. They annotated videos of human experts sharing knowledge with novices and found that the use of gestures was associated with both higher senses of warmth and competence, and smiling was associated with higher senses of warmth but lower competence. This was also found to extend to virtual agents in a follow-up study (Biancardi et al., 2017b). Bergmann et al. (2012) similarly emphasize how including gestures helps improve a user's first impressions of an agent when interacting with it for the first time. The researchers measured responses against the social cognition dimensions of warmth and competence. When gestures were present, feelings of competence increased, and likewise decreased when gestures were absent, aligning with the findings by Biancardi et al. (2017a).

Although these studies show how personality can be attributed to agents through their use of NVC, the appropriateness of those behaviors is equally important. In addition to their earlier findings, Cafaro et al. (2016) also showed how the appropriateness of an agent's behavior is crucial in forming first impressions of the agent. They explained how first impressions set a baseline for the human's expectations and is influential in deciding whether or not they should continue to interact with the agent. Through the use of both verbal and nonverbal behaviors (including proxemics, gaze, and smiling), an agent can express more relational capabilities such as showing empathy and friendliness (Cafaro et al., 2016). In their study, Cafaro et al. found that the use of these appropriate behaviors resulted in increased time spent with the agent, with greater reported values of user engagement and satisfaction. In addition, Kistler et al. (2012) further emphasized the importance of appropriate behaviors in an agent when they showed how people felt that agents that displayed nonverbal behavior aligning with their own cultural expectations were more appropriate.

5.3. Achieving a sense of copresence

Social presence, or copresence, is a sense of intimacy and immediacy between people (Short et al., 1976). As virtual agents strive to be social actors, it is vital that they provide a sense of copresence to the user. Research has focused on how both agent appearance and agent behavior create a sense of presence.

For example, Bailenson et al. (2005) studied how agent copresence is affected by appearance and behavior with regards to head movements. They compared how users' ratings of copresence were affected by both an agent's realism of motion and realism of appearance. They emphasize that the two are strongly related. Participants felt that for the least human-like agents, the least realistic motions were appropriate. Likewise, for the more human-like agents, more realistic motions were appropriate. They found that a mismatch between appearance and motion would lead to lower ratings and sense of copresence.

Even different degrees of agency can affect presence. To illustrate, Nowak and Biocca (2003) compared different agents in a virtual environment, varying their agency (such as having a speech-only agent) and anthropomorphism (robotic vs. human-like agent). They argue that an agent's appearance alone can increase a user's sense of copresence, just through the use of embodiment. In their study, people would naturally assume an agent as anthropomorphic unless the agent presented evidence to the contrary. For example, an agent without a visual appearance felt more humanlike than one with an obviously robotic appearance. However, Guadagno et al. (2007) provide an interesting nuance to this argument. They establish that higher levels of *behavioral* realism (not just a realistic appearance) produce higher levels of social presence. These findings show the importance of an agent's behavior (not just appearance or linguistic capabilities) and how users' incoming beliefs, perceptions, and expectations play a large role as well.

The way an agent responds to and interacts with users can also affect presence. An agent that provides realistic nonverbal feedback to a user's actions feels more copresent. Garau et al. (2005) studied how people perceived agents based on their movements and responsiveness with regards to proxemic behavior. For example, a responsive agent would look at the user when the user walked close to it in a VR environment. The researchers' goal was to see if agents would be treated as social entities. They asked participants to walk through a virtual environment that with a virtual agent in it, without explicitly asking them to interact with the agent. They found that users reported a higher sense of personal contact and copresence when interacting with the agents that would recognize the user's behavior and respond accordingly. Agents that spoke led users to engage with them more often. The non-engaging, unresponsive agents felt "ghostlike" as their behavior was not influenced by the user's actions. Behavior itself can affect how users socially perceive an agent even without speech. To quote the authors, "*On some level people can respond to agents as social actors even in the absence of two-way verbal interaction.*" This example points to the power of NVC in human-agent interactions.

5.4. Increasing rapport, trust, and empathy

Related to the sense of social presence is the concept of rapport. Rapport refers to the coordination and relationship between individuals and is related to a mutual sense of trust. In humans, rapport is strongly related to nonverbal behaviors, including expressions of positivity (like smiling), shared attention, and coordination (Tickle-Degnen & Rosenthal, 1990). Maintaining rapport with a user is important for agents in therapy/healthcare applications (DeVault et al., 2014; Gratch et al., 2006; Kang et al., 2012; Nguyen & Masthoff, 2009), but also teaching (Baylor & Kim, 2008, 2009; Rickel & Johnson, 1999, 2000) and collaboration when completing tasks (Andrist et al., 2017).

The Rapport Agent by Gratch et al. (2006) is a prime example of an agent that leverages NVC to increase rapport.

In a lab study using the Rapport Agent, Gratch, Wang, Gerten et al. (2007) showed how differing agent responsiveness affects human behavior and feelings of rapport. They tested different conditions: a face-to-face condition where two participants spoke directly with each other, a mediated condition where the "listener" participant was represented by an avatar instead, a responsive agent that displayed nonverbal listening behavior based on perceived speech and head movements from the user, and a prerecorded avatar that moved, but not in response to any user input. Surprisingly, Gratch et al. found that the responsive agent led to the highest amount of engagement and that many participants believed they were interacting with an avatar representing a human rather than an autonomous agent. Their results showed how the use of NVC can greatly improve the effectiveness of a virtual agent.

As with the agents described in earlier sections (Buschmeier & Kopp, 2018; Cassell & Thorisson, 1999; Garau et al., 2005), Gratch et al. (2006) found that responsiveness and feedback is key to an effective agent interaction. They emphasize that the contingency of feedback, not just the prevalence of behavioral cues and animation is important for creating feelings of rapport between a human and an agent. A responsive agent may even be better at creating feelings of rapport than a human (Gratch, Wang, Okhmatovskaia et al., 2007). Kang et al. (2008) continued these studies by looking at users' feelings of shyness and self-performance when interacting with contingent agents. They found that social anxiety decreased with an increase in the contingency of the agent's nonverbal behaviors. Greater anxiety led to decreased feelings of rapport and worse performance with the non-contingent agent. Huang et al. (2011) later improved on the Rapport Agent by using collected data of people interacting with prerecorded videos and implementing their specific behaviors into the agent. The agent also incorporates partial mirroring of the other's behavior. They found that this agent improved feelings of rapport and was better overall in terms of naturalness, turn-taking, etc.

Related to rapport is the notion of trust. As an example, Cowell and Stanney (2003) focused on how nonverbal behavior can create trust and credibility in virtual agent interactions. They designed an agent that could portray both trusting (e.g., increased eye contact) and non-trusting (e.g., gaze aversions, negative facial expressions) behaviors. Users presented with the trusting agent rated the agent as more credible and were more satisfied with the interaction than the non-trusting agent, with which they rated as less trustworthy. Thus, it is important to use proper nonverbal behaviors for an agent that needs the user to trust them.

Mirroring and mimicry can also increase the trustworthiness and persuasiveness of an agent, as shown by Bailenson and Yee (2005). Bailenson and Yee (2005) created an agent that would mimic a person's head movements as the agent delivered a persuasive message. The mimicking agent was more effective than one that did not mimic. However, when the mimicry was too obvious, people realized that the agent was copying them (Bailenson et al., 2008). The detection of mimicry decreased ratings of trust and friendliness. Their findings indicate that there may be a threshold to mirroring, balancing the potential benefits with the possibility that the behavior could backfire.

The effects of mirroring and trust may also depend on context. Verberne et al. (2013) evaluated mimicry in two scenarios: one that was more competence-based (trusting that the agent could do the job) and one that was more relational (trusting that the agent's intentions were pure). In the competence-based scenario, participants had to choose if they would trust the navigational skills of an agent to their own. In the relational-based scenario, participants had to choose whether or not to trust an agent with investing their money. For both scenarios, Verberne et al. compared the use of a head-motion mimicking agent vs a non-mimicking agent. The results from the experiment showed that although people liked and trusted the mimicking agent overall, the effects were more pronounced for the competence-based scenario than the relational one. Their results imply that, although mirroring behaviors can increase trust, the effects may be tempered by the context and experiences a user has with the agent.

An agent that displays empathetic emotions can also increase senses of trust and empathy. For example, Nguyen and Masthoff (2009) developed an agent that intervenes to alleviate a user's negative mood. When evaluating their system, they found that an agent that displayed empathic expressions led to higher positive ratings and likability from users. They establish that a human-like representation affords empathic capabilities and increases the sense of emotional intelligence in an agent. Likewise, Kang et al. (2012) showed how the use of specific nonverbal behaviors in a virtual counselor increased user perceptions of intimacy. A virtual counselor would require self-disclosure from people, thus requiring a degree of intimacy. Kang et al. first studied natural human behavior to identify eye gazes, head nods, head shakes, tilts, pauses, and smiles, which they then implemented into an agent. Participants that saw the agent expressing these emotional behaviors attributed a higher level of intimacy to the interaction. Even head movements alone (without any facial features/expressions) are able to increase trust and induce more disclosure of information from people (Bailenson et al., 2006), showing just how important NVC is at establishing trust.

5.5. Enhancing the efficacy of collaboration and learning

NVC also plays a large role in agents that teach and agents that collaborate with people to accomplish tasks. Research in these areas mainly focuses on studying how the inclusion of NVC improves memory recall and task performance. In humans, recall that gaze is an efficient way to coordinate attention between individuals (Brennan et al., 2008). Andrist et al. (2017) showed how gaze can be also used for coordinating actions in a virtual agent. They created a model of gaze behavior based on prior literature and implemented these behaviors into an agent that cooperated with users to build sandwiches. The researchers evaluated their agent in a user study, asking participants to collaborate with the agent to complete sandwich-building tasks (Figure 3). Andrist et al. found that an agent that responded to and produced gaze

resulted in a faster task completion time. Additionally, the agent was rated higher in terms of cognitive ability and competence, also resulting in higher amounts of shared gaze and mutual gaze. These results show how an agent's behavior can affect both users' performance and perception of the agent.

For pedagogical/teaching agents, NVC is often used to add richness and emphasize concepts when delivering information (Andrist et al., 2012a; Baylor & Kim, 2008, 2009; Bergmann & Macedonia, 2013; Mayer & DaPra, 2012). For instance, Andrist et al. (2012a) created a presenter agent that used referential gaze to increase information recall in users. They also showed how affiliative (i.e., mutual) gaze increased feelings of connectedness, but failed to show an increase in recall performance. Similarly, Baylor and Kim (2008) looked at the interplay between an agent's nonverbal behaviors and its teaching style. They found that the perception of behavior differed depending on the way a course was taught. For teaching attitudinal and persuasive information, facial expressions were valued more than deictic gestures, which were valued more in procedural, linear lecturing. This was followed up (Baylor & Kim, 2009) with a deeper analysis of gestures and teaching. The authors theorize that because expressions and gestures were both types of visual information, they may have interfered with users' working memory (cognitive load theory, Chandler & Sweller, 1991). When inappropriate, nonverbal cues can even evoke negative responses (Baylor & Kim, 2009); thus, agent behaviors must match the intended application and content to avoid detrimental or non-effects.

This theme is echoed by other research. Bergmann and Macedonia (2013) looked at the use of iconic gestures (those that illustrate words by painting a nonverbal picture) in agents. People were asked to learn foreign words from an agent; the researchers measured their learning performance and memory recall of the words. They found that including iconic gestures resulted in better performance than a control with no gestures. Surprisingly, Bergmann and Macedonia found that for long-term recall, the agent was more effective than a human when teaching higher-performing students (i.e., those who scored well on a short-term recall test). However, a human was better for teaching low-performing students. They speculate that the agent's behavior may have contributed additional cognitive load, which distracted the low-performing students, leading to a lessened effect. This finding emphasizes how important it is for an agent's behaviors to appear natural and appropriate.

Although gesture has been established to be beneficial for learning in humans (Cook & Goldin-Meadow, 2006; Goldin-Meadow & Alibali, 2013; Gorham, 1988; Novack & Goldin-Meadow, 2015), its effects are inconsistent in agents, with studies often unable to find a significant effect of NVC on learning outcomes (Baylor et al., 2003; Buisine et al., 2004; Buisine & Martin, 2007; Frechette & Moreno, 2010; Krämer et al., 2007; Wang & Gratch, 2009). To illustrate, Wang and Gratch (2009) used the Rapport Agent in a sexual harassment training context. The Rapport Agent would take users through a course and then they were asked to retell the information they learned. The researchers saw that including immediate feedback in the agent (in the form of nodding when spoken to, or mimicking the gaze of the speaker) helped increase feelings of rapport and helpfulness. The agent also increased subjective reports of self-efficacy (which may in turn help with learning) but was not able to significantly affect recall performance.

To further study the effects of gesture in agents, Buisine et al. (2004) looked at different gesturing strategies for agents when teaching or presenting information. Specifically, they looked at *redundant* (communicates same information as speech), *complementary* (communicates additional information alongside speech), and *speech-specialization* (gestures not intended to convey task-specific information; e.g., touching one's face) gestures and evaluated users' subjective impressions and ability to recall information taught by an agent. They found that gestures did not enable users to recall information any better, although users did rate the redundant and

Figure 3. A user interacting with an agent that employs gaze cues to help coordinate and direct the user to the intended target in a collaborative sandwich-building game. (Image from Andrist et al., 2017).

complementary gesturing agents higher. Essentially, there was only a difference in likability, not actual performance. A later study (Buisine & Martin, 2007) adjusted the behaviors to fix negatively perceived gestures from the prior study. This time, they found that speech-redundant gestures resulted in significantly higher recall of verbal information and also had higher ratings of quality and expressiveness. Multimodal redundancy may improve a user's social perception of an agent, viewing it as more likable and with a positive personality.

However, when designing pedagogical agents, likability is not as critical as an agent's ability to teach. In a comparison of agents that gestured against agents that did not, Krämer et al. (2003) found a surprising result. Users rated agents that used gesture as significantly more entertaining but less helpful than agents which did not gesture. They speculated that the gestures may have been too pronounced and therefore caused an adverse effect in users, echoing the cognitive load issues described earlier. It seems likely that the context of the agents may also have played a role. The agents were designed to provide a human-like interface that controlled a TV/VCR system. Users may have seen the use of gestures as too unnecessary, given the limited conversation needed to operate the functions of a VCR.

Frechette and Moreno (2010) also conducted a study to look at learning performance when instructed by a virtual agent. In their study, they looked at information recall and comprehension when learning about planetary systems. They compared agents that displayed combinations of affective facial expressions and cognitive thinking gestures against those that did not. They also compared a version of the teaching system with no agent. They found no significant differences between the agents, with the exception that affective nonverbals (such as facial expressions) resulted in lower performance than just a static, non-animated agent. This supports the criticism that agent NVC may serve as a distractor or may be counterproductive and not help learning outcomes. This may also be related to the appropriateness of the behaviors: affective behaviors may not have been proper for this type of teaching style and thus accrued more cognitive load, affecting the results. Overall, research has been inconsistent in determining the effectiveness of NVC in teaching agents.

6. Developing a virtual agent that uses NVC

In order to understand the challenges with creating a virtual agent that uses NVC, we must also understand how an agent is developed, in terms of the technical challenges that are inherent to such a complex system. Based on our survey, we identified common themes among many of the virtual agent systems described above that underlie the implementation of NVC in agents. In this section, we describe the general components that go into a virtual agent that uses NVC and how the individual subcomponents, or modules, are developed and integrated together. Additionally, we detail the methods used to understand and author agent behaviors, comparing the benefits and challenges of each approach.

6.1. What makes up a virtual agent?

To understand the design of virtual agent systems, we must first look at how dialogue systems are designed, or the forerunner to the embodied conversational agent. Spoken dialogue systems have traditionally followed an architecture that incorporates six components, or modules, that divide the system into logical parts (Jurafsky & Martin, 2000). First and foremost, these include input modules both to recognize words spoken by a user (speech recognition) and to parse the intent of the speech (natural language understanding). From there, a system needs to keep track of the current state of the conversation and plan the next course of action a system should take (dialogue manager). The dialogue manager may also need to perform specific actions by interfacing with another application (task manager). Following this comes the output modules to plan what a system should say next (natural language generation) and, finally, convert the text back into audible words for the user to hear (text-to-speech synthesis).

For the most part, virtual agents have followed a similar architecture. This is mainly due to their history as extensions of dialogue systems and their still prevalent need for speech. Virtual agent architectures include the modules listed above with the addition of nonverbal behavior modules to support the added display of an avatar. These may include modules to handle animation of the avatar's behaviors (e.g., producing gestures or displaying facial expressions) as well as modules for perceiving and recognizing any nonverbal input from the user.

A typical virtual agent architecture is best described by the pattern laid out by Matsuyama et al. (2016) in the recent development of an agent called SARA. They condense the architecture into a simple pipeline consisting of three main phases: *understanding, reasoning*, and *generation* (Figure 4). Below, we go deeper into each phase, describing the submodules involved in each phase and how agents have incorporated them.

6.1.1. Understanding phase

The *understanding* phase deals with converting the raw input from the user into a clean semantic form (such as the pertinent information in speech or the intent of a gesture, etc.) that the agent can easily parse. First, recognition modules use sensors to obtain input/information from the user (Jurafsky & Martin, 2000; Kopp et al., 2006; Matsuyama et al., 2016). For instance, a speech recognition module listens for a user's utterance from a microphone and converts it into text. Similarly, a facial expression recognition module tracks faces through a camera and uses computer vision to identify distinct expressions. After recognition, semantic understanding modules parse the inputs to infer the user's intent behind each utterance and nonverbal behavior. The utterance text is parsed by a natural language understanding module to extract the user's communicative intent. Likewise, recognized facial expressions are used to infer the user's current mood.

To start, developing NVC recognition modules is difficult (Kopp et al., 2004). Recognizing and interpreting user behavior is a large research effort on its own, and often intersects with the fields of machine learning, computer

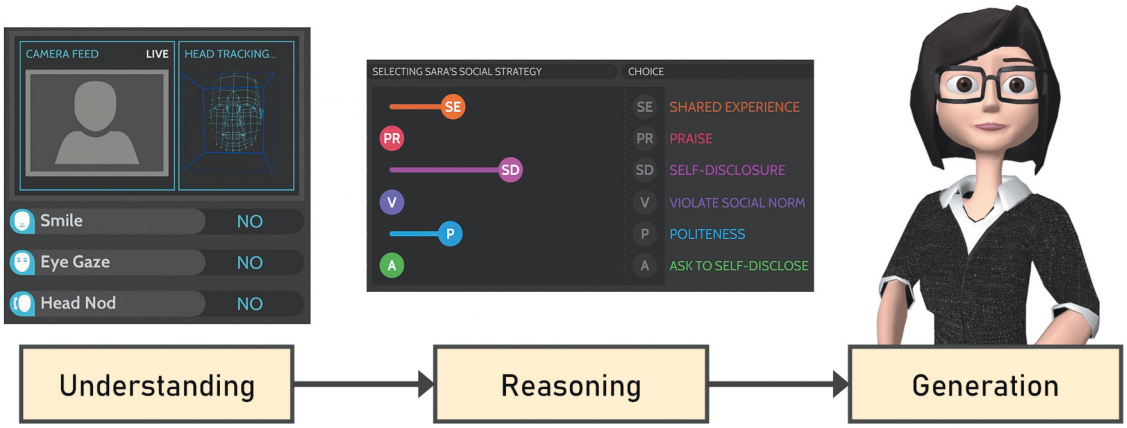

Figure 4. SARA agent and the three phases that make up her architecture, going from understanding (perceiving the user and interpreting their input) to reasoning (deciding on conversational strategy based on those inputs) and finally to generation (production of speech and NVC in an animated avatar) (Images modified from Matsuyama et al., 2016) Images courtesy Justine Cassell, Carnegie Mellon University.

vision, natural language processing and similar areas of expertise (Gratch et al., 2002). The difficulty lies in the challenge of, first, recognizing raw input from the user and, second, mapping those inputs to semantic intents (Kopp et al., 2004). Both input recognition and intent parsing modules require training machine learning classifiers (such as decision trees or support vector machines) or relying on heuristic models derived from psychological and communication literature.

Fortunately, for input recognition, researchers have utilized off-the-shelf components and existing recognition systems to help simplify parts of this process. As an example, Anderson et al. (2013) used the Microsoft Kinect sensor to capture human gestures, gaze, and facial expression. They developed an interview coach that used recognition to detect detrimental social cues during an interview. Researchers have also relied on the use of wearable sensors (Feese et al., 2012; Terven et al., 2016; Thórisson, 1997), motion sensing cameras (Cassell et al., 1999; Morency et al., 2006), gaze trackers (DeVault et al., 2014; Thórisson, 1997), facial expression libraries (Burleson et al., 2004; DeVault et al., 2014), and motion capture systems (Anderson et al., 2013; Lance & Marsella, 2009) to fill the need for recognition.

Although these systems convert user input (raw speech/motion) into standardized and manageable forms (text, gesture description, features, etc.), the issue remains on how to process this data and extract the user's semantic intent. It is one thing to understand the words that are being spoken, but another to actually comprehend them. In dialogue systems, this is the job of the natural language understanding module (Jurafsky & Martin, 2000). Virtual agent systems that use NVC need a similar understanding module, but for nonverbal behaviors.

Interpreting nonverbal behaviors requires a model that maps input behaviors to intents, either derived from prior work or mined using collected data. As an example, for the interview coach mentioned above, Anderson et al. (2013) trained a Bayesian network on the multimodal input in order to classify social cues. The agent would detect the user's affective state as well, progressing with the interview when appropriate cues were detected. Another example is one by Wang et al. (2020), which featured a model that estimated users' impressions of an agent by monitoring their facial expressions and leveraging this information to change how users felt about the agent. Other agents recognize gaze behaviors through complex models derived from psychological research (Andrist et al., 2017; Huang et al., 2011) or trained using data from user studies (Morency et al., 2006). Bevacqua et al. (2010) used a combination of trained models and heuristic rules to determine the state of the user and the conversation. Their rules were based on previous perceptual studies of human-human communication, which allowed them to accurately determine what a user's intent was based on their behaviors.

Oftentimes, these recognized inputs (speech, expression, gesture, etc.) may not only be parsed individually, but also combined to determine complex intents. For instance, the SARA agent (Matsuyama et al., 2016) uses a combination of utterance, facial expressions, gaze, and other auditory and visual cues as input to a classifier that infers the user's conversational strategy. Through the classifier, SARA estimates the current level of rapport with the user. She then strategizes how to respond with appropriate verbal and nonverbal behavior to increase the level of rapport. Similarly, the SimSensei agent (DeVault et al., 2014) tracks a user's facial expressions, gaze, and fidgeting motions. By combining these inputs in the MultiSense system (Gratch et al., 2013), the agent is able to estimate the user's level of anxiety and distress. This complex understanding of human behavior allows the agent to be more effective as a virtual therapist. Burleson et al. (2004) created a complex "inference engine" that takes in data from a wide variety of sensors to accurately determine a user's affective state. Another example is the engagement modeling by Dermouche and Pelachaud (2019), who combined facial expressions, head gestures, and gaze to interpret the level of engagement from the user. Although multimodal

recognition adds complexity to an agent, it allows the agent to interact in a much more human-like manner; human-human communication is often multimodal in nature (Quek et al., 2002).

6.1.2. Reasoning phase

The extracted intents produced by the understanding phase are fed into the reasoning phase to determine what the agent should do next. The *reasoning* phase deals with taking in the intents derived from the understanding phase and deciding what the agent should do next in the interaction. This typically involves any dialogue management or state-keeping in the system and represents the bulk of an agent's intelligence. In a dialogue system, this is mainly the job of the dialogue manager module (Jurafsky & Martin, 2000). Although often not explicitly called a "dialogue manager" due to the addition of NVC, virtual agents utilize similar modules for deciding an agent's next move in response to the user.

An agent's actions may be determined by simple rule-based decisions (Cassell et al., 1999; Lee et al., 2007), a finite state machine (Matsuyama et al., 2016), or more complex AI networks (Cafaro et al., 2016; Pelachaud, 2017), among other methods. An agent's intelligence may even incorporate task and domain knowledge, allowing it to teach specific subjects (Rickel & Johnson, 1999, 2000) or answer questions (Cassell et al., 1999; Thórisson, 1997). At a simple level, an agent may decide to directly proceed with dialogue (Anderson et al., 2013), show empathy (Nguyen & Masthoff, 2009), or even nod to indicate attention (Huang et al., 2011; Lee et al., 2007). Often, these actions are a direct result of user input. On a higher level, an agent may have its own goals or even personality that governs its actions, regardless of the user's direct input.

For instance, the Max agent (Becker et al., 2004) introduces the concept of mood. Max features an internal model of overall mood and current emotion which can change over the course of an interaction. These moods are modeled after the pleasure-arousal-dominance (PAD) scale (Mehrabian, 1969), which is a way to describe temperament based on those three dimensions. Max's mood dictates how the agent should interpret specific statements and respond. For example, he may get bored of an interaction or even become annoyed if the user says something offensive. This will cause him to start making different facial expressions and more assertive gestures. If the user continues their behavior, Max may even get angry and decide to leave the conversation until the emotion system "cools off."

This notion of a simulated mental state is not unique to Max. Several researchers have adopted and advocated for the use of an internal agent state (in the form of goals, attitude, etc.) that drives behavior (Becker et al., 2004; Gratch et al., 2002; Lee et al., 2007; Poggi et al., 2000, 2005; Traum et al., 2008). One notable example is the negotiation agent created by Traum et al. (2008). The agent must solve problems and gain a user's trust, all while managing the conversation. The agent's internal goal is to persuade; thus, it tries to actively understand the beliefs and goals of others while employing different strategies for negotiation based on changes in their behavior.

Pelachaud et al. (2002) argue that agents should have an internal model of belief, desire, and intent (BDI), presenting their Greta agent as an example. Greta utilizes a complex dynamic belief network to represent BDI. The network governs when to show emotion, when to hide them, and changes emotional state over time. All verbal and nonverbal behaviors are a result of the agent's internal state of mind RPPC03. (Rosis et al., 2003). Poggi et al. (2005) argue that including a BDI model is necessary for agents to act natural and believable. They emphasize that the production of both speech and gesture should come from a common intent. These concepts align with the theory that, in humans, gesture is directly tied to the production of thought (McNeill, 1992).

Thus, NVC can even be considered a window into the internal state of an agent. That is the principle Lee et al. (2007) considered when creating the Rickel Gaze Model, a novel way for agents to subtly communicate their intent. Recall that gaze can be used to give feedback, direct attention, and even pass/hold the speaking floor. The Rickel Gaze Model drives an agent's gaze behaviors (e.g., looking away) for conveying its internal state (e.g., planning an utterance). These behaviors allow for giving realistic feedback that is valuable for agent-human communication.

6.1.3. Generation phase

The *generation* phase deals with converting the agent's next action or intent into real audio and visual output that the user can perceive. This can basically be thought of as the reverse of the understanding phase; here, we start with an intent and convert it back into speech and nonverbal behaviors. For dialogue systems, this consists of two modules, natural language generation, which takes in an agent's intent and generates the corresponding utterance text, followed by text-to-speech, which takes in the text and produces actual audible speech (Jurafsky & Martin, 2000). A virtual agent system must do the same but include the production of NVC in addition to speech.

Instead of a language generation module, agents utilize a behavior generator that takes in the agent's intent and determines the type of NVC to produce, and when to produce it (Kipp et al., 2010). The module typically produces a description of the behavior, with parameters that dictate how they are to be performed. The earliest agents used the XML format to store this information, but replacements tailored to virtual agent applications have since been developed (Badler et al., 2000; Heloir & Kipp, 2009; Kipp et al., 2010; Kopp et al., 2006; Kranstedt et al., 2002). The abstract description of the behavior is then passed to a behavior "realizer," a module that governs the kinematics, animation, and synchronization of behaviors in the agent's avatar (Kipp et al., 2010).

Behavior generation can be as simple as manually scripting intended NVC into the dialogue. For instance, Kranstedt et al. (2002) introduced a markup language called MURML (Multimodal Utterance Representation Markup Language) that allows text to be tagged with gestures. The tags describe the specific hand and arm motions that comprises each gesture and when to perform them in alignment with speech. However, manually authoring dialogue scripts and annotating them with gestures is time consuming and may not

represent natural behavior, as they are subject to the author's best judgment. Instead, researchers have developed complex systems that automatically generate intended behaviors with or without accompanying text (Bergmann & Kopp, 2009; Cassell et al., 1994, 2004; Ravenet et al., 2018; Salem & Earle, 2000).

The BodyChat system by Cassell and Vilhjálmsson (1999) was an early system designed to automatically animate avatars based on different conversational behaviors. Although their system was designed for human-controlled avatars, not autonomous agents, it was capable of animating gaze, gesture, and facial expressions to match the user's inputted text. They later introduced the Behavior Expression Animation Toolkit (BEAT) (Cassell et al., 2004). The BEAT system takes in text, parses it to determine clauses, objects, actions, etc., and suggests different nonverbal behaviors that would be appropriate. The capabilities range from simple "beat" gestures that sync with each spoken word to performing iconic gestures that illustrate different actions.

These methods rely on parsing text to determine the underlying communicative intent and then applying nonverbal behaviors that correspond to those behaviors. Lee and Marsella (2006) analyzed video clips of human interactions to create a technique that automatically generates behaviors based on speech. Their technique involved first parsing the given text to determine the type of utterance, such as an affirmation or interjection. Based on the type, they would then choose a nonverbal behavior based on behavior rules that they identified through the video clips. For example, the text "I suppose" would be classified as a "possibility," which would produce head nods and raised eyebrows.

Other techniques generate gestures that convey spatial information based on context. For instance, Bergmann and Kopp (2009) created an agent that automatically generates gestures for giving directions, based on the spatial relationship of the agent to locations in real life. Another system, developed by Ravenet et al. (2018), would create appropriate metaphoric gestures (i.e. gestures used to try and depict abstract ideas) from user speech. Their technique focused on mapping words to physical representations and gestures (Figure 5). If an agent said the word "rise," it would make a rising gesture to accompany it. If it talked about a table, that would result in a broad "wiping" gesture representing the table's surface.

These systems focus on generating appropriate behaviors; however, this only determines *what* gestures, gaze, etc. to perform. The latter part of the generation phase involves taking the generated behaviors and then animating them in an avatar. This is the job of the behavior realizer, the NVC analogue of text-to-speech. Realizing these behaviors can be as simple as playing back a prerecorded animation, as was done in some early agents (Badler et al., 1999; Cassell et al., 1994). However, this method lacks a lot of the nuance and dynamics of real NVC. A capable behavior realizer must also plan the individual motor actions of an avatar and also synchronize them with speech and other actions (Kopp et al., 2006). Doing so may involve interpreting abstract behaviors into actual motor control expressions (Heloir & Kipp, 2009), planning arm and hand articulation for gestures (Kopp et al., 2004), and tweaking the motion to express different intents (Badler et al., 1999), all to animate nonverbal behaviors in an agent.

6.2. Defining behaviors in virtual agents

One main challenge behind creating a virtual agent that uses NVC is first identifying proper nonverbal behaviors (Allwood et al., 2007). This entails understanding what the nonverbal signals mean, and how they are used in natural human-human communication. Understanding behaviors is a necessary first step that governs the entire design and development of an agent. Designing appropriate behaviors (both verbal and nonverbal) is crucial for an agent's success. How realistic an agent appears and behaves can impact how users accept the agent and interact with it (Koda & Maes, 1996; Parise et al., 1999; Wagner et al., 2006). Similarly, the concept of *believability* also plays an important part.

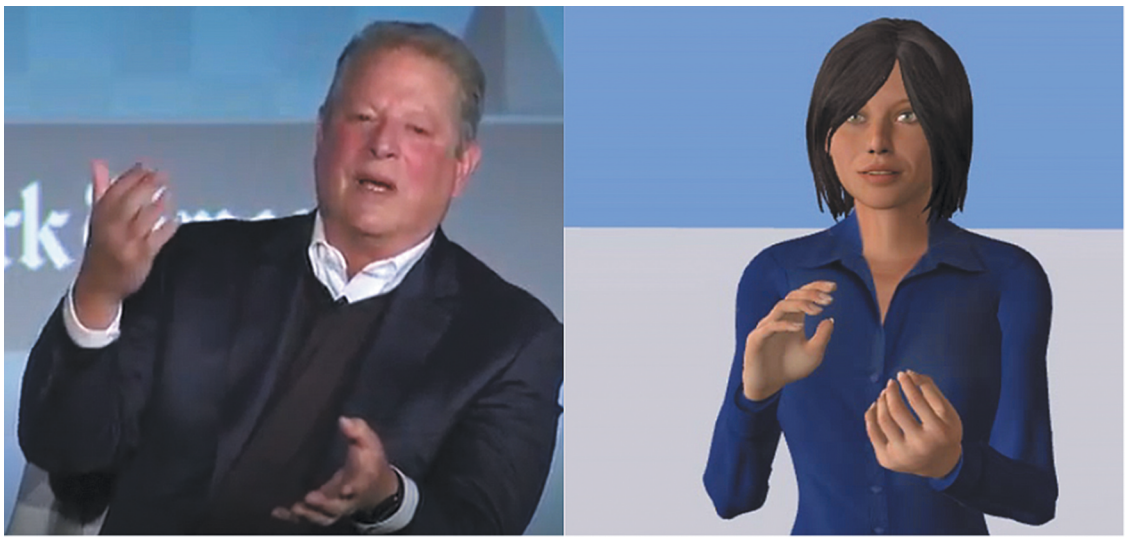

Figure 5. A person (left) describing two ideas by gesturing separately with each hand; when given the same speech text, a virtual agent (right) automatically produces similar gestures. (Image from Ravenet et al., 2018).

Believability differs from realism in that, while realism is related to the visual appearance and movement of an agent, believability relates to the sense of an agent as a real, living being; that it exists beyond its visual appearance (Pelachaud et al., 2002). An agent's use of NVC must match what users expect from natural human-human NVC in order to be believable (Burleson et al., 2004; McRorie et al., 2012).

However, manually authoring proper NVC behaviors is challenging, limited, and often does not accurately reflect real human motions (André & Pelachaud, 2010; Badler et al., 1999). Thus, researchers have relied on different methods for identifying and implementing NVC in agents. In our review, we describe three such approaches for defining NVC in virtual agents. These are the model-driven, data-driven, and crowd-driven methods. We summarize each method below and discuss the differences between each approach.

6.2.1. Model-driven

One method, the "model-driven" approach as defined by Rehm et al. (2007), implements nonverbal behaviors based on derived models of behavior reported in prior work. A lot of early agents were designed using this approach (Becker et al., 2004; Cassell et al., 1999; Isbister et al., 2000; Pelachaud, 2005a; Rickel & Johnson, 2000). The main advantage of the model-driven approach is that it is relatively simple. It relies on existing models of behavior, typically from psychological or social science literature. These articles are based off observation of human behavior and present models of how gestures are used and the purposes they serve, such as those described in Section 3.

Using existing models allows behaviors to be condensed into simple rules. For example, an agent should wave if a user makes a greeting, or nod to signify understanding when the user makes a declarative statement (Rickel & Johnson, 2000). Noma et al. (2000), who created one of the earliest agents that featured real-time generated NVC, first studied prior work on presentations and public speaking to understand how people use gestures when speaking. They then derived a set of rules specifying when an agent should use a specific gesture during its speech. Likewise, the agents by Kenny et al. (2007) also interact using similarly derived rules from the social sciences, behaving appropriately based on each agent's intent.

Other researchers have gone beyond simple rule-based interaction, instead combining several existing models to understand how NVC should be articulated. Pelachaud (2009b) combined the findings from four perceptual studies on gesture expressivity to create gestures that exhibited different emotions and personality. They also did the same for facial expressions, modeling how expressions change over time to convey emotion (Pelachaud, 2009a). Pejsa et al. (2015) even went as far as to create a mathematical model of how the eyes, head, and torso move to convey gaze based on literature from both psychology and human kinematics.

Prior work provides good insight into the human processes that recognize and produce behavior. As a result, the model-driven approach is a good heuristic method for quickly producing natural behaviors in agents. However, although model-driven approaches are typically easier to implement, they are less nuanced than real human motion and may not be suitable for all applications (Rehm & André, 2008).

6.2.2. Data-driven

One limitation of relying heavily on derived models is that, although they provide a good high-level description of what behavior to perform in each circumstance (rules), they fail to capture the details or fine movements of actual behavior. This is where the second method, a "data-" or "corpus-driven" approach, comes in. A data-driven method requires first collecting a large amount of recorded footage of human interactions, which are then analyzed to identify key behaviors that an agent should perform. The recorded data can be of human-human interactions, or human-agent interactions (such as live recorded interactions with an agent or from a Wizard of Oz study). The annotated data may be analyzed manually using statistical methods (Allwood et al., 2007; Biancardi et al., 2017a, 2017b. Lance & Marsella, 2009) or machine learning models (Bilakhia et al., 2013; Dermouche & Pelachaud, 2019; Kopp et al., 2004; Matsuyama et al., 2016; Morency et al., 2006; Ochs, Libermann et al., 2017).

For example, the SARA agent mentioned in the previous section (Matsuyama et al., 2016) is a primarily data-driven agent. For SARA's social reasoner module, Matsuyama et al. trained classifiers from collected data to determine a user's conversational strategy and level of rapport. To determine NVC used in conversation, Allwood et al. (2007) analyzed interactions between pairs of strangers conversing with each other. They recorded videos of 30 pairs, totaling up to over an hour of footage. The researchers fully coded the videos, noting changes in posture, facial expressions, and gaze. As a result, they were able to model how agents should present nonverbal feedback as a user speaks. Similarly, Foster and Oberlander (2007) created a system to generate head and eyebrow movements based on a corpus of annotated videos. Additionally, they conducted an evaluation study and found that users preferred data-driven generation than a more generalized model-driven approach. Their finding shows that data-driven methods captures more of the nuance in natural NVC.

One of the advantages of a data-driven method is that real data can represent the full range of nuance present in real human NVC (which a "distilled" model of behavior fails to capture). This approach is useful when there are no existing studies or models for the specific behaviors that an agent is trying to employ. For example, the smoking cessation coach by Grolleman et al. (2006) required them to first analyze the behaviors of a real coach. Although prior work modeled similar situations with interviewing and coaching, the researchers needed to understand the specific nonverbal strategies that a smoking coach employs. Similarly, Gratch et al. (2013) needed to understand the nonverbal signs associated with distress during clinical interviews. As part of this process, they first collected a large dataset or interview recordings. They then analyzed the data to find correlations between distress and the nonverbal behaviors of gaze, facial expression, and gesture, for later feeding recognition models.

Data-driven methods also allow researchers to capture specific uses of NVC that are more subtle or have not been previously studied (Pelachaud & Poggi, 2002). Lance and Marsella (2009) present such a case. Their goal was to understand how gaze can be used to express emotion; however, they

did not find any prior literature that mapped the two. Thus, they conducted a study to collect data of human-human interactions, annotated the data, and used it to statistically derive a mapping of gaze to emotion. Even more complex is the problem of producing a sequence of different behaviors over time to express a specific attitude or stance. By applying temporal sequence mining techniques on annotated data, both Chollet et al. (2014) and Janssoone et al. (2016) were able to create models to generate different sequences of gestures, facial expressions, etc. to convey an agent's attitude to the user.

The data-driven method has also been combined with the aforementioned model-driven method to form a hybrid approach. When creating agents that represent different cultures, Rehm and André (2008) primarily relied on a theoretical model of culture by Hofstede et al. (2010) but collected and analyzed their own data to fill in the gaps. Their data supplemented the broad categories from the model with real data. Likewise, Andrist (2013) and Andrist, Mutlu et al. (2013) use a hybrid method to implement gaze in agents. They combine high-level rules from theoretical models and actual data to emulate the low-level subtlety of gaze shifts and aversions. Similarly, Lance and Marsella (2010), Lance & Marsella (2008) improved on their earlier data-driven approach by incorporating models from literature on nonverbal acting and affect, allowing them to create a robust mapping of gaze to different emotions.

6.2.3. Crowd-driven

Despite the benefits of data-driven approaches, they require data collection, annotation, and analysis, which are time-consuming (Rehm & André, 2008). In our literature review, we identified a third method that has become more common in recent years. The "crowd-driven" approach relies on the power of the crowd to assist with annotation and authoring of NVC behaviors. Ravenet et al. (2013) employed this method to generate gesture, gaze, and facial expressions that conveyed different interpersonal attitudes. They created a web interface that allowed users to select different nonverbal behaviors and change their attributes (such as a forceful arm gesture or a subtle head tilt). The researchers presented users with prompts asking them to create a gesture that would be appropriate for a given scenario (e.g., showing a submissive attitude). They collected over 900 submissions, which they were able to use to directly train a Bayesian network to replicate the behaviors.

Similarly, Ochs, Pelachaud et al. (2017) noted that smiles may communicate information based on context and subtle differences in the smile. Rather than record data of people smiling (and requiring annotation), the researchers utilized crowdsourcing to collect a large number of manually created smiles. Using a similar web interface (Ravenet et al., 2013), they presented users with a written scenario and asked them to change smile attributes (size, symmetry, speed, etc.) to express the feelings in the scenario. They used the collected data to quickly develop a decision tree model for expressing different emotions using smiles alone. Ochs et al. later evaluated and showed the effectiveness of their crowdsourcing approach to generating NVC from user perception.

Using a crowd-driven method may even be more effective than a data-driven approach, due to the use of a large number of people rather than a small number of coders. Dermouche and Pelachaud (2018) used crowdsourcing to evaluate their existing generation model. The original model mapped NVC to the characteristics of friendliness and dominance. However, in their crowdsourcing study, they were able to find how the behaviors affected specific perceived attitude and personality traits. Huang et al. (2010, 2011) conducted a study where they asked participants to watch a video of a person telling a story. Participants were asked to pretend the person in the video was conversing with them. The participants pressed a button anytime during the video they felt like feedback (in the form of a head nod or affirmative "uh-huh") would be appropriate. Essentially, Huang et al. were able to simplify and crowdsource the annotation of nonverbal behaviors. They then were able to identify the ideal times for an agent to display feedback showing that they are listening. They evaluated their model and found that their new method produced higher feelings of rapport and believability than the previous annotation method. Thus, crowd-driven approaches may be better at generating behaviors that are more believable and perceivable, all while reducing the time and effort needed for annotation.

7. Discussion of challenges and themes

Based on our review, we discuss the main challenges with creating virtual agents that use NVC. We identified three main themes in the literature: 1) the difficulty in creating large-scale annotated datasets, 2) a diversity in agent implementation and need for standardization, and 3) the goal of creating appropriate NVC.

7.1. The need for annotation

One of the main challenges of creating a virtual agent that uses NVC revolves around understanding how NVC works in human-human interactions. A key requirement is that agent behaviors must be realistic and believable (Koda & Maes, 1996; Parise et al., 1999; Pelachaud et al., 2002; Wagner et al., 2006). As a result, there has been a large emphasis on first understanding how NVC is naturally used in humans, and then implementing those behaviors into agents.

The current authoring approaches that we detailed in the previous section have different strengths and weaknesses. The model-driven approach allows for easy authoring of behaviors based on existing literature. However, the technique lacks the nuance of NVC as it condenses human behavior into generalized rules and tendencies. In contrast, the more comprehensive data-driven approach involves collecting data/recording of humans and analyzing them to understand how nonverbal behavior works. The collected data can even be used for training both recognition and generation models. The use of real recorded data has the benefit of including any subtle expressions or motions that a real human would use.

The main drawback to this approach is the need for annotation. Videos need to be annotated to label the salient features and behaviors that the agent must emulate. Annotation involves identifying regions of video when specific behaviors

occur, alongside transcription of speech. To assist in this process, researchers have created annotation tools (Baur et al., 2013; Kipp, 2001; Wang et al., 2018; Wittenburg et al., 2006) that allow for easier labeling of multimodal behaviors in video. These tools even feature different levels of automation to help offload some of the burden of annotation. Even with the use of specialized tools, data annotation can still be a time-consuming process, as human coders are needed to label each multimodal signal that occurs in the data.

Thus, to mitigate the need for annotation, researchers have designed hybrid approaches that use existing models for governing general behavior along with using data to drive fine-grained motions (Andrist, 2013; Andrist, Mutlu et al., 2013; Rehm & André, 2008). Another approach is through crowd-driven methods, which seek to distribute the work. These approaches either delegate annotation to the crowd (Huang et al., 2010, 2011) or directly allow the crowd to author behaviors (Ochs, Pelachaud et al., 2017; Ravenet et al., 2013). Crowdsourcing can even evaluate how behaviors are perceived by people, ensuring that a large population interpret the signals correctly (Andrist et al., 2012b). It even enables models to be reevaluated for new effects (such as understanding how personality is affected by articulation) (Dermouche & Pelachaud, 2018). However, crowd-driven techniques are relatively new and still need to be refined and evaluated. Overall, authoring and annotation is still a difficult, time-consuming process that needs to be improved.

7.2. The call for standardization

The other main difficulty we identified through our literature review deals with the technical aspects of implementation, particularly the research field's need for standardization. With development of NVC-enabled agents being so difficult, researchers have focused on creating standardized interfaces and architectures that allow for researchers to share technologies (De Carolis et al., 2004; Gratch et al., 2002; Kopp et al., 2006). As a result, a number of different frameworks and languages exist for handling nonverbal behaviors in agents.

For example, the SAIBA architecture (Kopp et al., 2006) focuses on separating agents into three layers, intent planning, behavior planning, and realization planning (similar to the three phases of *understanding, reasoning*, and *generation*). In particular, the Greta agent and derivatives (Niewiadomski et al., 2009; Pelachaud, 2017; Poggi et al., 2005) use the SAIBA architecture. The goal of SAIBA is to separate the different layers of functionality and introduce standardized interfaces to drive the development of reusable components. However, despite the existence of SAIBA, researchers have proposed other frameworks and architectures that similarly separate agent functionality, but do so in different ways.

Another virtual agent framework, SEMAINE (Schröder, 2010), focuses on real-time multimodal feedback and has been used to study distinct characteristics of agent behavior (Bevacqua et al., 2010; McRorie et al., 2012; Ochs, Pelachaud et al., 2017). It differs from SAIBA in that the framework revolves around emotion and features tracking the current user and agent affective state. The Virtual Human Toolkit (Hartholt et al., 2013), used by SimSensei (DeVault et al., 2014) and SARA (Matsuyama et al., 2016), aims to make it easy to create agents for different purposes and places an emphasis on perceiving user behavior by supporting a number of different sensors. On the other hand, agents such as Max (Becker et al., 2004) and Rea (Cassell et al., 1999) rely on their own architectures, tailored specifically for their applications.

To pass information between components, researchers have also created markup languages that describe nonverbal behaviors. For instance, FML (Heylen et al., 2008) and PML (Scherer et al., 2012) describe the nonverbal input perceived from the user, in terms of intents and abstract behaviors. For marking up text with different nonverbal behaviors, researchers have introduced APML (De Carolis et al., 2004), EmotionML (Schröder et al., 2011), MURML (Kranstedt et al., 2002), and BML (Kopp et al., 2006), all differing slightly in the level of granularity they represent and the different types of NVC they support. In addition, researchers have also made attempts to merge languages, but have just created others in the process, as in the case of FML-APML (Mancini & Pelachaud, 2008). Out of all these markup languages, in our review, we saw that BML was the most adopted language (as part of the SAIBA architecture (Kopp et al., 2006)), with use in the Rapport Agent (Huang et al., 2011), SimSensei (DeVault et al., 2014), SARA (Matsuyama et al., 2016), and a number of behavior generation systems (Kenny et al., 2007; Kipp et al., 2010; Lee & Marsella, 2006). However, the other languages were only briefly mentioned in the literature, or limited to a few authors.

The issue with both the markup languages and the architectures is that they end up becoming competing standards, as emphasized by Schröder when describing the SEMAINE framework (Schröder, 2010). Schröder points out a need for the virtual agent community to agree on identifying challenges in integrating technologies and subsequently define specifications and standards for these technologies. Although the idea of standardization is good, and researchers have made efforts to consolidate agent NVC technologies, there is still a wide range of diversity in the approach, featureset, and design of these implementations.

7.3. The goal of appropriateness and consistency

For the most part, virtual agents have been able to emulate human NVC with similar effects. The ability to express emotion appears to translate well to virtual agents (Becker et al., 2004; Nguyen & Masthoff, 2009; Poggi et al., 2005). Agents that gesture with varying articulation can effectively convey personality and attitude (Biancardi et al., 2017a; McRorie et al., 2012; Neff et al., 2010). Similarly, the strategic use of NVC also increases perceptions of warmth and likability in an agent (Bergmann et al., 2012; Cafaro et al., 2016). Studies have also shown how an agent displaying listening feedback and nonverbal mirroring can increase a user's sense of rapport and trust (Bailenson & Yee, 2005; Cowell & Stanney, 2003; Gratch et al., 2006; Gratch, Wang, Gerten et al., 2007; Huang et al., 2011). Even some of the more functional aspects of NVC translate to agents, such as managing conversations (Andrist, Leite et al., 2013; Cassell et al., 1999) and directing attention (Andrist et al., 2017; Buschmeier & Kopp, 2018).

However, despite these successes, there were still several instances where studies failed to show a significant effect of NVC on users' outcomes. In particular, most of the issues we saw involved the use of NVC for teaching, and may have been due to a lack of appropriateness and consistency in agent behaviors.

To recap, research on natural human communication has predominantly shown the use of gesture, gaze, and attention to increase the effectiveness of learning (Cook & Goldin-Meadow, 2006; Goldin-Meadow & Alibali, 2013; Gorham, 1988; Novack & Goldin-Meadow, 2015). Gesturing is essential to the communication of thoughts (McNeill, 1992); thus, when used in a learning context, they increase the memorability and understanding of the presented content (Goldin-Meadow & Alibali, 2013). In agents, Andrist et al. (2012a) showed how deictic gaze can improve learning through its ability to refer to objects in a natural and unobtrusive way. Baylor and Kim (2008, 2009) saw that NVC increased learning in specific scenarios: gestures were effective when learning procedural tasks and facial expressions were effective for attitudinal/persuasive presentations. Likewise, Bergmann and Macedonia (2013) found that the use of iconic gestures helped with teaching vocabulary words.

In contrast, Wang and Gratch (2009) showed how nonverbal feedback increased a user's subjective feelings of self-efficacy, but not actual recall performance. On a similar note, Buisine et al. (2004) found that users liked agents that gestured when delivering information, but the researchers also found no effect on learning performance. In a study by Krämer et al. (2007), users reported that the inclusion of gestures made an agent feel more entertaining but subjectively less helpful. Frechette and Moreno (2010) echo this thought; in their own study, users rated nonverbal behaviors as distracting.

While the use of NVC did not negatively impact the outcomes of these studies, their inability to show a significant effect may be due to the appropriateness of the NVC used. The studies by Baylor and Kim (2008, 2009), Bergmann and Macedonia (2013), and Andrist et al. (2012a) all support this possibility. An agent's nonverbal behaviors need to match the type of content that it is presenting (Baylor & Kim, 2008). For instance, displaying a range of emotions through facial expression would not be ideal when the goal is to inform the user about the solar system (Frechette & Moreno, 2010). Buisine et al. (2004), Buisine & Martin (2007) provide an excellent example of the need for appropriate behaviors. After failing to find any effect of gesture on learning performance in their initial study (Buisine et al., 2004), the authors replaced the negatively received gestures with more appropriate and natural ones. When they later reevaluated their system with the new gestures, they reported that speech-redundant gestures helped with the recall of information (Buisine & Martin, 2007), which aligns with prior work (Cook & Goldin-Meadow, 2006; Goldin-Meadow & Alibali, 2013). A similar effect was reported by Berry et al. (2005), who found that an agent led to poor memory performance when its facial expressions were inconsistent with the content; their findings point to consistency as another possible contributor to agent success. This notion echoes the thoughts of other researchers who emphasize that a believable agent must behave consistently with coherent speech, appearance, and nonverbal behaviors (McRorie et al., 2012; Niewiadomski et al., 2010; Pelachaud et al., 2002). A decreased sense of believability can also negatively affect how users accept and interact with the agent (Koda & Maes, 1996; Parise et al., 1999; Wagner et al., 2006).

In summary, when an agent's behaviors clash with the rest of its presentation, the user may experience additional cognitive load from attempting to make sense of the conflicting signals (Baylor & Kim, 2009). In a learning context, the cognition required may already be high, so any inappropriate use of NVC will reduce an agent's efficacy (Bergmann & Macedonia, 2013). The varability of results also relates back to the need for better methods of understanding human behavior and creating models of NVC, so that researchers can create agents with appropriate and consistent behaviors.

8. Conclusion

In this paper, we presented a literature review of virtual agents that use nonverbal communication (NVC). We established how NVC plays an important role in natural human communication and highlighted the use of NVC in virtual agents to emulate the expressivity and multimodality that NVC affords. We also elaborated on the different functions and outcomes of using NVC in agents and how NVC is implemented and defined in those agents. Based on our review, we identified challenges and themes across the literature. For the most part, virtual agents have been successful in using NVC to communicate conversational intent and attention, and also improve user perceptions of friendliness, trust, rapport, etc. However, research showed inconclusive results on the effectiveness of NVC in agents for areas such as teaching and training, despite a clear benefit in humans. The likely cause is a lack of appropriateness in virtual agent behavior. Furthermore, this may be due to the underlying difficulty in defining proper behaviors in virtual agents, due to the need for annotation, analysis, and lack of full standardization for NVC in the virtual agent community. These themes point to a continued need for research into virtual agent NVC and creating agents that are more lifelike, further closing the gap between human-human and human-agent interactions.

Funding

This work is partially supported by the National Science Foundation under grant award [#IIS-1750840]. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect these agencies' views.

ORCID

Isaac Wang http://orcid.org/0000-0003-0613-6112

References

Alibali, M. W. (2005, December). Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. *Spatial Cognition and Computation*, *5*(4), 307–331. https://doi.org/10.1207/s15427633scc0504_2

Allbeck, J. M., & Badler, N. I. (2001). Towards behavioral consistency in animated agents. In N. Magnenat-Thalmann & D. Thalmann (Eds.), *Deformable avatars* (pp. 191–205). Springer US. https://doi.org/10.1007/978-0-306-47002-8%5F17

Allwood, J., Kopp, S., Grammer, K., Ahlsén, E., Oberzaucher, E., & Koppensteiner, M. (2007, December). The analysis of embodied communicative feedback in multimodal corpora: A prerequisite for behavior simulation. *Language Resources and Evaluation*, *41*(3), 255–272. https://doi.org/10.1007/s10579-007-9056-2

Anderson, K., André, E., Baur, T., Bernardini, S., Chollet, M., Chryssafidou, E., Damian, I., Ennis, C., Egges, A., Gebhard, P., Jones, H., Ochs, M., Pelachaud, C., Porayska-Pomsta, K., Rizzo, P., & Sabouret, N. 2013. The tardis framework: Intelligent virtual agents for social coaching in job interviews. In D. Reidsma, H. Katayose, & A. Nijholt (Eds.), *Advances in computer entertainment* (pp. 476–491). Springer International Publishing.

André, E., & Pelachaud, C. (2010). Interacting with embodied conversational agents. In F. Chen (Ed.), *Speech technology: Theory and applications* (pp. 123–149). Springer US. https://doi.org/10.1007/978-0-387-73819-2%5F8

Andrist, S. (2013). Controllable models of gaze behavior for virtual agents and humanlike robots. *Proceedings of the 15th ACM on International Conference on Multimodal Interaction* (pp. 333–336). ACM. https://doi.org/10.1145/2522848.2532194

Andrist, S., Mutlu, B., & Gleicher, M. (2013). Conversational gaze aversion for virtual agents. In R. Aylett, B. Krenn, C. Pelachaud, & H. Shimodaira (Eds.), *Intelligent virtual agents* (pp. 249–262). Springer.

Andrist, S., Gleicher, M., & Mutlu, B. (2017). Looking coordinated: Bidirectional gaze mechanisms for collaborative interaction with virtual characters. *Proceedings of the 2017 chi Conference on Human Factors in Computing Systems* (pp. 2571–2582). ACM. https://doi.org/10.1145/3025453.3026033

Andrist, S., Leite, I., & Lehman, J. (2013). Fun and fair: Influencing turn-taking in a multi-party game with a virtual agent. *Proceedings of the 12th International Conference on Interaction Design and Children* (pp. 352–355). ACM. https://doi.org/10.1145/2485760.2485800

Andrist, S., Pejsa, T., Mutlu, B., & Gleicher, M. (2012a). Designing effective gaze mechanisms for virtual agents. *Proceedings of the Sigchi Conference on Human Factors in Computing Systems* (pp. 705–714). ACM. https://doi.org/10.1145/2207676.2207777

Andrist, S., Pejsa, T., Mutlu, B., & Gleicher, M. (2012b). A head-eye coordination model for animating gaze shifts of virtual characters. *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction* (pp. 4:1–4:6). Santa Monica, California: ACM. https://doi.org/10.1145/2401836.2401840

Argyle, M. (1988). *Bodily communication*. Methuen.

Argyle, M., & Dean, J. (1965). Eye-contact, distance and affiliation. *Sociometry*, *28*(3), 289–304. https://doi.org/10.2307/2786027

Badler, N. I., Bindiganavale, R., Allbeck, J., Schuler, W., Zhao, L., & Palmer, M. (2000). Parameterized action representation for virtual human agents. *Embodied Conversational Agents* (pp. 256–284). MIT Press. http://dl.acm.org/citation.cfm?id=371552.371567

Badler, N. I., Chi, D. M., & Chopra-Khullar, S. (1999, May). Virtual human animation based on movement observation and cognitive behavior models. *Proceedings Computer Animation 1999* (pp. 128–137).

Bailenson, J. N., Beall, A. C., & Blascovich, J. (2002, December). Gaze and task performance in shared virtual environments. *The Journal of Visualization and Computer Animation*, *13*(5), 313–320. https://doi.org/10.1002/vis.297

Bailenson, J. N., Swinth, K., Hoyt, C., Persky, S., Dimov, A., & Blascovich, J. (2005, August). The independent and interactive effects of embodied-agent appearance and behavior on self- report, cognitive, and behavioral markers of copresence in immersive virtual environments. *Presence: Teleoperators and Virtual Environments*, *14*(4), 379–393. https://doi.org/10.1162/105474605774785235

Bailenson, J. N., & Yee, N. (2005, October). Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological Science*, *16*(10), 814–819. https://doi.org/10.1111/j.1467-9280.2005.01619.x

Bailenson, J. N., Yee, N., Merget, D., & Schroeder, R. (2006, August). The effect of behavioral realism and form realism of real-time avatar faces on verbal disclosure, nonverbal disclosure, emotion recognition, and copresence in dyadic interaction. *Presence: Teleoperators and Virtual Environments*, *15*(4), 359–372. https://doi.org/10.1162/pres.15.4.359

Bailenson, J. N., Yee, N., Patel, K., & Beall, A. C. (2008, January). Detecting digital chameleons. *Computers in Human Behavior*, *24*(1), 66–87. https://doi.org/10.1016/j.chb.2007.01.015

Bartlett, M. S., Hager, J. C., Ekman, P., & Sejnowski, T. J. (1999, March). Measuring facial expressions by computer image analysis. *Psychophysiology*, *36*(2), 253–263. https://doi.org/10.1017/S0048577299971664

Baur, T., Damian, I., Lingenfelser, F., Wagner, J., & André, E. (2013). Nova: Automated analysis of nonverbal signals in social interactions. In A. A. Salah, H. Hung, O. Aran, & H. Gunes (Eds.), *Human behavior understanding* (pp. 160–171). Springer International Publishing.

Bavelas, J. B., & Chovil, N. (2006). Nonverbal and verbal communication: Hand gestures and facial displays as part of language use in face-to-face dialogue. In V. Manusov & M. L. Patterson (Eds.), *The sage handbook of nonverbal communication* (pp. 97–116). SAGE Publications, Inc. http://sk.sagepub.com/reference/hdbk%5Fnonverbalcomm/n6.xml

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, *52*(3), 566–580. https://doi.org/10.1111/j.1460-2466.2002.tb02562.x

Baylor, A., Ryu, J., & Shen, E. (2003). The effects of pedagogical agent voice and animation on learning, motivation and perceived persona. *Association for the Advancement of Computing in Education (AACE)*, 452–458.

Baylor, A. L., & Kim, S. (2008). The effects of agent nonverbal communication on procedural and attitudinal learning outcomes. In H. Prendinger, J. Lester, & M. Ishizuka (Eds.), *Intelligent virtual agents* (pp. 208–214). Springer Berlin Heidelberg.

Baylor, A. L. (2009, December). Promoting motivation with virtual agents and avatars: Role of visual presence and appearance. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3559–3565. https://doi.org/10.1098/rstb.2009.0148

Baylor, A. L. (2011, April). The design of motivational agents and avatars. *Educational Technology Research and Development*, *59*(2), 291–300. https://doi.org/10.1007/s11423-011-9196-3

Baylor, A. L., & Kim, S. (2009, March). Designing nonverbal communication for pedagogical agents: When less is more. *Computers in Human Behavior*, *25*(2), 450–457. https://doi.org/10.1016/j.chb.2008.10.008

Becker, C., Kopp, S., & Wachsmuth, I. (2004). Simulating the emotion dynamics of a multi- modal conversational agent. In E. André, L. Dybkjær, W. Minker, & P. Heisterkamp (Eds.), *Affective dialogue systems* (pp. 154–165). Springer Berlin Heidelberg.

Becker, C., Kopp, S., & Wachsmuth, I. (2007). Why emotions should be integrated into conversational agents. *Conversational Informatics* (pp. 49–67). John Wiley & Sons, Ltd. https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470512470.ch3

Beebe, S. A. (1976). Effects of eye contact, posture and vocal inflection upon credibility and comprehension. *Australian SCAN: Journal of Human Communication*, *7*(8), 57–70.

Bergmann, K. (2006). Verbal or visual? How information is distributed across speech and gesture in spatial dialog. *Proceedings of Brandial 2006, the 10th Workshop on the Semantics and Pragmatics of Dialogue* (pp. 90–97).

Bergmann, K., Eyssel, F., & Kopp, S. (2012). A second chance to make a first impression? How appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time. In

Y. Nakano, M. Neff, A. Paiva, & M. Walker (Eds.), *Intelligent virtual agents* (pp. 126–138). Springer Berlin Heidelberg.

Bergmann, K., & Macedonia, M. (2013). A virtual agent as vocabulary trainer: Iconic gestures help to improve learners' memory performance. In R. Aylett, B. Krenn, C. Pelachaud, & H. Shimodaira (Eds.), *Intelligent virtual agents* (pp. 139–148). Springer Berlin Heidelberg.

Bergmann, K., & Kopp, S. (2009). Increasing the expressiveness of virtual agents: Autonomous generation of speech and gesture for spatial description tasks. *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 1* (pp. 361–368). International Foundation for Autonomous Agents and Multiagent Systems. http://dl.acm.org/citation.cfm?id=1558013.1558062

Berry, D. C., Butler, L. T., & De Rosis, F. (2005, September). Evaluating a realistic agent in an advice-giving task. *International Journal of Human-Computer Studies*, *63*(3), 304–327. https://doi.org/10.1016/j.ijhcs.2005.03.006

Bevacqua, E., Raouzaiou, A., Peters, C., Caridakis, G., Karpouzis, K., Pelachaud, C., & Mancini, M. (2006). Multimodal sensing, interpretation and copying of movements by a virtual agent. In E. André, L. Dybkjær, W. Minker, H. Neumann, & M. Weber (Eds.), *Perception and interactive technologies* (pp. 164–174). Springer Berlin Heidelberg.

Bevacqua, E., Pammi, S., Hyniewska, S. J., Schröder, M., & Pelachaud, C. (2010). Multimodal backchannels for embodied conversational agents. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (Eds.), *Intelligent virtual agents* (pp. 194–200). Springer Berlin Heidelberg.

Biancardi, B., Cafaro, A., & Pelachaud, C. (2017a). Analyzing first impressions of warmth and competence from observable nonverbal cues in expert-novice interactions. *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (pp. 341–349). ACM. https://doi.org/10.1145/3136755.3136779

Biancardi, B., Cafaro, A., & Pelachaud, C. (2017b). Could a virtual agent be warm and competent? Investigating user's impressions of agent's non-verbal behaviours. *Proceedings of the 1st ACM Sigchi International Workshop on Investigating Social Interactions with Artificial Agents* (pp. 22–24). ACM. https://doi.org/10.1145/3139491.3139498

Bilakhia, S., Petridis, S., & Pantic, M. (2013, September). Audiovisual detection of behavioural mimicry. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (pp. 123–128). IEEE.

Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B., & Zelinsky, G. J. (2008, March). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, *106*(3), 1465–1477. https://doi.org/10.1016/j.cognition.2007.05.012

Buisine, S., Abrilian, S., & Martin, J.-C. (2004). Evaluation of multimodal behaviour of embodied agents. In Z. Ruttkay & C. Pelachaud (Eds.), *From brows to trust: Evaluating embodied conversational agents* (pp. 217–238). Springer Netherlands. https://doi.org/10.1007/1-4020-2730-3%5F8

Buisine, S., & Martin, J.-C. (2007, July). The effects of speech–gesture cooperation in animated agents' behavior in multimedia presentations. *Interacting with Computers*, *19*(4), 484–493. https://doi.org/10.1016/j.intcom.2007.04.002

Burleson, W., Picard, R. W., Perlin, K., & Lippincott, J. (2004). A platform for affective agent research. *Workshop on Empathetic Agents, International Conference on Autonomous Agents and Multiagent Systems, Columbia University, New York, NY* (Vol. 2). Citeseer.

Buschmeier, H., & Kopp, S. (2018). Communicative listener feedback in human-agent interaction: Artificial speakers need to be attentive and adaptive. *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems* (pp. 1213–1221). International Foundation for Autonomous Agents and Multiagent Systems. http://dl.acm.org/citation.cfm?id=3237383.3237880

Cafaro, A., Vilhjálmsson, H. H., Bickmore, T., Heylen, D., Jóhannsdóttir, K. R., & Valgarosson, G. S. (2012). First impressions: Users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In Y. Nakano, M. Neff, A. Paiva, & M. Walker (Eds.), *Intelligent virtual agents* (pp. 67–80). Springer Berlin Heidelberg.

Cafaro, A., Vilhjálmsson, H. H., & Bickmore, T. (2016, August). First impressions in human–agent virtual encounters. *ACM Transactions on Computer-Human Interaction*, *23*(4), 24: 1–24:40. https://doi.org/10.1145/2940325

Caridakis, G., Raouzaiou, A., Bevacqua, E., Mancini, M., Karpouzis, K., Malatesta, L., & Pelachaud, C. (2007, December). Virtual agent multimodal mimicry of humans. *Language Resources and Evaluation*, *41*(3), 367–388. https://doi.org/10.1007/s10579-007-9057-1

Cassell, J. (2001, December). Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. *Embodied Conversational Agents* (pp. 1–27). MIT Press.

Cassell, J., Vilhjálmsson, H. H., & Bickmore, T. (2004). Beat: The behavior expression animation toolkit. In H. Prendinger & M. Ishizuka (Eds.), *Life-like characters: Tools, affective functions, and applications* (pp. 163–185). Springer. https://doi.org/10.1007/978-3-662-08373-4%5F8

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., & Yan, H. (1999). Embodiment in conversational interfaces: Rea. *Proceedings of the Sigchi Conference on Human Factors in Computing Systems* (pp. 520–527). ACM. https://doi.org/10.1145/302979.303150

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., … Stone, M. (1994). Animated conversation: Rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques* (pp. 413–420). ACM. https://doi.org/10.1145/192161.192272

Cassell, J., & Thorisson, K. R. (1999, May). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, *13*(4–5), 519–538. https://doi.org/10.1080/088395199117360

Cassell, J., & Vilhjálmsson, H. (1999, March). Fully embodied conversational avatars: Making communicative behaviors autonomous. *Autonomous Agents and Multi-Agent Systems*, *2*(1), 45–64. https://doi.org/10.1023/A:1010027123541

Castellano, G., Mancini, M., Peters, C., & McOwan, P. W. (2012, May). Expressive copying behavior for social agents: A perceptual analysis. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *42*(3), 776–783. https://doi.org/10.1109/TSMCA.2011.2172415

Chandler, P., & Sweller, J. (1991, December). Cognitive load theory and the format of instruction. *Cognition and Instruction*, *8*(4), 293–332. https://doi.org/10.1207/s1532690xci0804_2

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*(6), 893. https://doi.org/10.1037/0022-3514.76.6.893

Chollet, M., Ochs, M., & Pelachaud, C. (2014). From non-verbal signals sequence mining to bayesian networks for interpersonal attitudes expression. In T. Bickmore, S. Marsella, & C. Sidner (Eds.), *Intelligent virtual agents* (pp. 120–133). Springer International Publishing.

Clarebout, G., Elen, J., Johnson, W. L., & Shaw, E. (2002). Animated pedagogical agents: An opportunity to be grasped? *Journal of Educational Multimedia and Hypermedia*, *11*(3), 267–286.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. Resnick, B. M. John, & S. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 13–1991). American Psychological Association.

Clark, H. H., & Wilkes-Gibbs, D. (1986, February). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7

Clavel, C., Plessier, J., Martin, J.-C., Ach, L., & Morel, B. (2009). Combining facial and postural expressions of emotions in a virtual character. In Z. Ruttkay, M. Kipp, A. Nijholt, & H. H. Vilhjálmsson (Eds.), *Intelligent virtual agents* (pp. 287–300). Springer Berlin Heidelberg.

Cook, M. (1977). Gaze and mutual gaze in social encounters: How long–and when–we look others "in the eye" is one of the main signals in nonverbal communication. *American Scientist*, *65*(3), 328–333.

Cook, S. W., & Goldin-Meadow, S. (2006, April). The role of gesture in learning: Do children use their hands to change their minds? *Journal*

of Cognition and Development, *7*(2), 211–232. https://doi.org/10.1207/s15327647jcd0702_4

Cowell, A. J., & Stanney, K. M. (2003). Embodiment and interaction guidelines for designing credible, trustworthy embodied conversational agents. In T. Rist, R. S. Aylett, D. Ballin, & J. Rickel (Eds.), *Intelligent virtual agents* (pp. 301–309). Springer Berlin Heidelberg.

Dael, N., Mortillaro, M., & Scherer, K. R. (2012). Emotion expression in body action and posture. *Emotion*, *12*(5), 1085–1101. https://doi.org/10.1037/a0025737

De Carolis, B., Pelachaud, C., Poggi, I., & Steedman, M. (2004). Apml, a markup language for believable behavior generation. In H. Prendinger & M. Ishizuka (Eds.), *Life-like characters: Tools, affective functions, and applications* (pp. 65–85). Springer. https://doi.org/10.1007/978-3-662-08373-4%5F4

Dermouche, S., & Pelachaud, C. (2018). Attitude modeling for virtual character based on temporal sequence mining: Extraction and evaluation. *Proceedings of the 5th International Conference on Movement and Computing* (pp. 23:1–23:8). ACM. https://doi.org/10.1145/3212721.3212806

Dermouche, S., & Pelachaud, C. (2019). Engagement modeling in dyadic interaction. *2019 International Conference on Multimodal Interaction* (pp. 440–445). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3340555.3353765

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., & Morency, L.P. (2014). Simsensei kiosk: A virtual human interviewer for healthcare decision support. *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems* (pp. 1061–1068). International Foundation for Autonomous Agents and Multiagent Systems. http://dl.acm.org/citation.cfm?id=2617388.2617415

Dillenbourg, P., & Traum, D. (2006, January). Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. *Journal of the Learning Sciences*, *15*(1), 121–151. https://doi.org/10.1207/s15327809jls1501_9

Duffy, K. A., & Chartrand, T. L. (2015, June). Mimicry: Causes and consequences. *Current Opinion in Behavioral Sciences*, *3*(6), 112–116. https://doi.org/10.1016/j.cobeha.2015.03.002

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, *23*(2), 283–292. https://doi.org/10.1037/h0033031

Duncan, S., & Niederehe, G. (1974, May). On signalling that it's your turn to speak. *Journal of Experimental Social Psychology*, *10*(3), 234–247. https://doi.org/10.1016/0022-1031(74)90070-5

Ekman, P. (1992, December). Are there basic emotions? *Psychological Review*, *99*(3), 550. https://doi.org/10.1037/0033-295X.99.3.550

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, *48*(4), 384–392. https://doi.org/10.1037/0003-066X.48.4.384

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., Scherer, K., Tomita, M., & Tzavaras, A. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, *53*(4), 712–717. https://doi.org/10.1037/0022-3514.53.4.712

Feese, S., Arnrich, B., Tröster, G., Meyer, B., & Jonas, K. (2012, September). Quantifying behavioral mimicry by automatic detection of nonverbal cues from body motion. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing* (pp. 520–525).

Foster, M. E., & Oberlander, J. (2007, December). Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Language Resources and Evaluation*, *41*(3), 305–323. https://doi.org/10.1007/s10579-007-9055-3

Frechette, C., & Moreno, R. (2010). The roles of animated pedagogical agents' presence and nonverbal communication in multimedia learning environments. *Journal of Media Psychology: Theories, Methods, and Applications*, *22*(2), 61–72. https://doi.org/10.1027/1864-1105/a000009

Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007, June). Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin*, *133*(4), 694. https://doi.org/10.1037/0033-2909.133.4.694

Fussell, S. R., Kraut, R. E., & Siegel, J. (2000). Coordination of communication: Effects of shared visual context on collaborative work. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work* (pp. 21–30). ACM. https://doi.org/10.1145/358916.358947

Fussell, S. R., Setlock, L. D., Yang, J., Ou, J., Mauer, E., & Kramer, A. D. I. (2004, September). Gestures over video streams to support remote collaboration on physical tasks. *Human–Computer Interaction*, *19*(3), 273–309. https://doi.org/10.1207/s15327051hci1903_3

Garau, M., Slater, M., Pertaub, D.-P., & Razzaque, S. (2005, February). The responses of people to virtual humans in an immersive virtual environment. *Presence: Teleoperators and Virtual Environments*, *14*(1), 104–116. https://doi.org/10.1162/1054746053890242

Gergle, D., Kraut, R. E., & Fussell, S. R. (2004). Action as language in a shared visual space. *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work* (pp. 487–496). ACM. https://doi.org/10.1145/1031607.1031687

Goldin-Meadow, S., & Alibali, M. W. (2013). Gesture's role in speaking, learning, and creating language. *Annual Review of Psychology*, *64*(1), 257–283. https://doi.org/10.1146/annurev-psych-113011-143802

Gorham, J. (1988, January). The relationship between verbal teacher immediacy behaviors and student learning. *Communication Education*, *37*(1), 40–53. https://doi.org/10.1080/03634528809378702

Grandhi, S. A., Joue, G., & Mittelberg, I. (2011). Understanding naturalness and intuitiveness in gesture production: Insights for touchless gestural interfaces. *Proceedings of the Sigchi Conference on Human Factors in Computing Systems* (pp. 821–824). ACM. https://doi.org/10.1145/1978942.1979061

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., Van Der Werf, R. J., & Morency, L.-P. (2006). Virtual rapport. In J. Gratch, M. Young, R. Aylett, D. Ballin, & P. Olivier (Eds.), *Intelligent virtual agents* (pp. 14–27). Springer.

Gratch, J., Wang, N., Okhmatovskaia, A., Lamothe, F., Morales, M., Van Der Werf, R. J., & Morency, L.-P. (2007). Can virtual humans be more engaging than real ones? In J. A. Jacko (Ed.), *Human-computer interaction. hci intelligent multimodal interaction environments* (pp. 286–297). Springer Berlin Heidelberg.

Gratch, J., Wang, N., Gerten, J., Fast, E., & Duffy, R. (2007). Creating rapport with virtual agents. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, & D. Pelé (Eds.), *Intelligent virtual agents* (pp. 125–138). Springer Berlin Heidelberg.

Gratch, J., Morency, L.-P., Scherer, S., Stratou, G., Boberg, J., Koenig, S., Adamson, T., & Rizzo, A. (2013). User-state sensing for virtual health agents and telehealth applications. *Studies in Health Technology and Informatics*, *184*(1), 151–157.

Gratch, J., Rickel, J., André, E., Cassell, J., Petajan, E., & Badler, N. (2002, July). Creating interactive virtual humans: Some assembly required. *IEEE Intelligent Systems*, *17*(4), 54–63. https://doi.org/10.1109/MIS.2002.1024753

Grolleman, J., Van Dijk, B., Nijholt, A., & Van Emst, A. (2006). Break the habit! Designing an e- therapy intervention using a virtual coach in aid of smoking cessation. In W. A. IJsselsteijn, Y. A. W. De Kort, C. Midden, B. Eggen, & E. Van Den Hoven (Eds.), *Persuasive technology* (pp. 133–141). Springer Berlin Heidelberg.

Guadagno, R. E., Blascovich, J., Bailenson, J. N., & Mccall, C. (2007, June). Virtual humans and persuasion: The effects of agency and behavioral realism. *Media Psychology*, *10*(1), 1–22. https://www.tandfonline.com/doi/full/10.1080/15213260701300865

Hall, E. T. (1966). *The hidden dimension*. Doubleday. (Google-Books-ID: TchK2tDnpkAC).

Hartholt, A., Traum, D., Marsella, S. C., Shapiro, A., Stratou, G., Leuski, A., … Gratch, J. (2013). All together now. In R. Aylett, B. Krenn, C. Pelachaud, & H. Shimodaira (Eds.), *Intelligent virtual agents* (pp. 368–381). Springer.

Hartmann, B., Mancini, M., & Pelachaud, C. (2006). Implementing expressive gesture synthesis for embodied conversational agents. In

S. Gibet, N. Courty, & J.-F. Kamp (Eds.), *Proceedings of the 6th international conference on gesture in human-computer interaction and simulation* (pp. 188–199). Springer Berlin Heidelberg. https://doi.org/10.1007/11678816%5F22

Hartmann, B., Mancini, M., Buisine, S., & Pelachaud, C. (2005). Design and evaluation of expressive gesture synthesis for embodied conversational agents. *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems* (pp. 1095–1096). New York, NY: ACM. Retrieved February 2, 2018, from https://doi.org/10.1145/1082473.1082640

Heidig, S., & Clarebout, G. (2011, January). Do pedagogical agents make a difference to student motivation and learning? *Educational Research Review*, *6*(1), 27–54. https://doi.org/10.1016/j.edurev.2010.07.004

Heloir, A., & Kipp, M. (2009). Embr – A realtime animation engine for interactive embodied agents. In Z. Ruttkay, M. Kipp, A. Nijholt, & H. H. Vilhjálmsson (Eds.), *Intelligent virtual agents* (pp. 393–404). Springer Berlin Heidelberg.

Heylen, D., Kopp, S., Marsella, S. C., Pelachaud, C., & Vilhjálmsson, H. (2008). The next step towards a function markup language. In H. Prendinger, J. Lester, & M. Ishizuka (Eds.), *Intelligent virtual agents* (pp. 270–280). Springer.

Hirsh, A. T., George, S. Z., & Robinson, M. E. (2009, May). Pain assessment and treatment disparities: A virtual human technology investigation. *Pain*, *143*(1), 106–113. https://doi.org/10.1016/j.pain.2009.02.005

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). *Cultures and organizations: Software of the mind* (3rd ed.). McGraw-Hill Education.

Horstmann, G. (2003). What do facial expressions convey: Feeling states, behavioral intentions, or actions requests? *Emotion*, *3*(2), 150–166. https://doi.org/10.1037/1528-3542.3.2.150

Huang, L., Morency, L.-P., & Gratch, J. (2011). Virtual rapport 2.0. In H. H. Vilhjálmsson, S. Kopp, S. Marsella, & K. R. Thórisson (Eds.), *Intelligent virtual agents* (pp. 68–79). Springer Berlin Heidelberg.

Huang, L., Morency, L.-P., & Gratch, J. (2010). Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior. *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1* (pp. 1265–1272). International Foundation for Autonomous Agents and Multiagent Systems. http://dl.acm.org/citation.cfm?id=1838206.1838371

Isbister, K., Nakanishi, H., Ishida, T., & Nass, C. (2000). Helper agent: Designing an assistant for human-human interaction in a virtual meeting space. *Proceedings of the Sigchi Conference on Human Factors in Computing Systems* (pp. 57–64). ACM. https://doi.org/10.1145/332040.332407

Jacob, C., Guéguen, N., Martin, A., & Boulbry, G. (2011, September). Retail salespeople's mimicry of customers: Effects on consumer behavior. *Journal of Retailing and Consumer Services*, *18*(5), 381–388. https://doi.org/10.1016/j.jretconser.2010.11.006

Janssoone, T., Clavel, C., Bailly, K., & Richard, G. (2016). Using temporal association rules for the synthesis of embodied conversational agents with a specific stance. In D. Traum, W. Swartout, P. Khooshabeh, S. Kopp, S. Scherer, & A. Leuski (Eds.), *Intelligent virtual agents* (pp. 175–189). Springer International Publishing.

Johnson, W. L., Rickel, J. W., & Lester, J. C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, *11*(1), 47–78.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (1st ed.). Prentice Hall PTR.

Kang, S.-H., Gratch, J., Sidner, C., Artstein, R., Huang, L., & Morency, L.-P. (2012). Towards building a virtual counselor: Modeling nonverbal behavior during intimate self-disclosure. *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1* (pp. 63–70). International Foundation for Autonomous Agents and Multi- agent Systems. http://dl.acm.org/citation.cfm?id=2343576.2343585

Kang, S.-H., Gratch, J., Wang, N., & Watt, J. H. (2008). Does the contingency of agents' nonverbal feedback affect users' social anxiety? *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1* (pp. 120–127). International Foundation for Autonomous Agents and Multiagent Systems. http://dl.acm.org/citation.cfm?id=1402383.1402405

Kendon, A. (1970, January). Movement coordination in social interaction: Some examples described. *Acta Psychologica*, *32*(2), 101–125. https://doi.org/10.1016/0001-6918(70)90094-6

Kendon, A. (1988). How gestures can become like words. In *Cross-cultural Perspectives in Nonverbal Communication* (pp. 131–141). Hogrefe.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Kenny, P., Hartholt, A., Gratch, J., Swartout, W., Traum, D., Marsella, S., & Piepol, D. (2007). Building interactive virtual humans for training environments. *Proceedings of I/ITSEC* (Vol. 174, pp. 911–916). NTSA.

Kipp, M. (2001). Anvil - a generic annotation tool for multimodal dialogue. *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)* (pp. 1367–1370). ISCA.

Kipp, M., Heloir, A., Schröder, M., & Gebhard, P. (2010). Realizing multimodal behavior. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (Eds.), *Intelligent virtual agents* (pp. 57–63). Springer Berlin Heidelberg.

Kistler, F., Endrass, B., Damian, I., Dang, C. T., & André, E. (2012, July). Natural interaction with culturally adaptive virtual characters. *Journal on Multimodal User Interfaces*, *6*(1), 39–47. https://doi.org/10.1007/s12193-011-0087-z

Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009, January). Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology*, *51*(1), 7–15. https://doi.org/10.1016/j.infsof.2008.09.009

Koda, T., & Maes, P. (1996, November). Agents with faces: The effect of personification. *Proceedings 5th IEEE International Workshop on Robot and human communication. Roman'96 tsukuba* (pp. 189–194). IEEE.

Kopp, S., Krenn, B., Marsella, S., Marshall, A. N., Pelachaud, C., Pirker, H., … Vilhjálmsson, H. (2006). Towards a common framework for multimodal generation: The behavior markup language. In J. Gratch, M. Young, R. Aylett, D. Ballin, & P. Olivier (Eds.), *Intelligent virtual agents* (pp. 205–217). Springer Berlin Heidelberg.

Kopp, S., Tepper, P., & Cassell, J. (2004). Towards integrated microplanning of language and iconic gesture for multimodal output. In *Proceedings of the 6th International Conference on Multimodal Interfaces* (pp. 97–104). ACM. https://doi.org/10.1145/1027933.1027952

Kopp, S., & Wachsmuth, I. (2004). Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds*, *15*(1), 39–52. https://doi.org/10.1002/cav.6

Krämer, N., Kopp, S., Becker-Asano, C., & Sommer, N. (2013, March). Smile and the world will smile with you–the effects of a virtual agent's smile on users' evaluation and behavior. *International Journal of Human-Computer Studies*, *71*(3), 335–349. https://doi.org/10.1016/j.ijhcs.2012.09.006

Krämer, N. C., Tietz, B., & Bente, G. (2003). Effects of embodied interface agents and their gestural activity. In T. Rist, R. S. Aylett, D. Ballin, & J. Rickel (Eds.), *Intelligent virtual agents* (pp. 292–300). Springer Berlin Heidelberg.

Krämer, N. C., Simons, N., & Kopp, S. (2007). The effects of an embodied conversational agent's nonverbal behavior on user's evaluation and behavioral mimicry. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, & D. Pelé (Eds.), *Intelligent virtual agents* (pp. 238–251). Springer Berlin Heidelberg.

Krämer, N. C., & Bente, G. (2010, March). Personalizing e-learning. the social effects of pedagogical agents. *Educational Psychology Review*, *22*(1), 71–87. https://doi.org/10.1007/s10648-010-9123-x

Kranstedt, A., Kopp, S., & Wachsmuth, I. (2002). Murml: A multimodal utterance representation markup language for conversational agents. *AAMAS'02 Workshop Embodied Conversational Agents - Let's Specify and Evaluate Them!*. ACM. https://pub.uni-bielefeld.de/publication/1857788

Kraut, R. E., Fussell, S. R., & Siegel, J. (2003, June). Visual information as a conversational resource in collaborative physical tasks. *Human–Computer Interaction*, *18*(1), 13–49. https://doi.org/10.1207/S15327051HCI1812_2

LaFrance, M. (1985, June). Postural mirroring and intergroup relations. *Personality & Social Psychology Bulletin*, *11*(2), 207–217. https://doi.org/10.1177/0146167285112008

Lafrance, M., & Broadbent, M. (1976, September). Group rapport: Posture sharing as a nonverbal indicator. *Group & Organization Studies*, *1*(3), 328–333. https://doi.org/10.1177/105960117600100307

Lakin, J. L., & Chartrand, T. L. (2003, July). Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science*, *14*(4), 334–339. https://doi.org/10.1111/1467-9280.14481

Lance, B., & Marsella, S. (2009, May). Glances, glares, and glowering: How should a virtual human express emotion through gaze? *Autonomous Agents and Multi-Agent Systems*, *20*(1), 50. https://doi.org/10.1007/s10458-009-9097-6

Lance, B., & Marsella, S. (2010, July). The expressive gaze model: Using gaze to express emotion. *IEEE Computer Graphics and Applications*, *30*(4), 62–73. https://doi.org/10.1109/MCG.2010.43

Lance, B. J., & Marsella, S. C. (2008). A model of gaze for the purpose of emotional expression in virtual embodied agents. *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1* (pp. 199–206). International Foundation for Autonomous Agents and Multiagent Systems. http://dl.acm.org/citation.cfm?id=1402383.1402415

Lee, J., & Marsella, S. (2006). Nonverbal behavior generator for embodied conversational agents. In J. Gratch, M. Young, R. Aylett, D. Ballin, & P. Olivier (Eds.), *Intelligent virtual agents* (pp. 243–255). Springer Berlin Heidelberg.

Lee, J., Marsella, S., Traum, D., Gratch, J., & Lance, B. (2007). The rickel gaze model: A window on the mind of a virtual human. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, & D. Pelé (Eds.), *Intelligent virtual agents* (pp. 296–303). Springer Berlin Heidelberg.

Mancini, M., & Pelachaud, C. (2008). The fml-apml language. *Proceedings of the Workshop on fml at aamas* (Vol. 8).

Matsuyama, Y., Bhardwaj, A., Zhao, R., Romeo, O., Akoju, S., & Cassell, J. (2016). Socially- aware animated intelligent personal assistant agent. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 224–227). Association for Computational Linguistics. http://aclweb.org/anthology/W16-3628

Mayer, R. E., & DaPra, C. S. (2012). An embodiment effect in computer-based learning with animated pedagogical agents. *Journal of Experimental Psychology. Applied*, *18*(3), 239–252. https://doi.org/10.1037/a0028616

McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., & Kaliouby, R. E. (2016). Affdex sdk: A cross-platform real-time multi-face expression recognition toolkit. *Proceedings of the 2016 chi Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 3723–3726). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2851581.2890247

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press. (Includes bibliographical references (pp. 393–407) and index.).

McRorie, M., Sneddon, I., McKeown, G., Bevacqua, E., De Sevin, E., & Pelachaud, C. (2012, July). Evaluation of four designed virtual agent personalities. *IEEE Transactions on Affective Computing*, *3*(3), 311–322. https://doi.org/10.1109/T-AFFC.2011.38

Mehrabian, A. (1969). Significance of posture and position in the communication of attitude and status relationships. *Psychological Bulletin*, *71*(5), 359–372. https://doi.org/10.1037/h0027349

Mehrabian, A. (1972). *Nonverbal communication*. Transaction Publishers.

Morency, L.-P., Christoudias, C. M., & Darrell, T. (2006). Recognizing gaze aversion gestures in embodied conversational discourse. *Proceedings of the 8th International Conference on Multimodal Interfaces* (pp. 287–294). ACM. https://doi.org/10.1145/1180995.1181051

Neff, M., Wang, Y., Abbott, R., & Walker, M. (2010). Evaluating the effect of gesture and language on personality perception in conversational agents. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (Eds.), *Intelligent virtual agents* (pp. 222–235). Springer Berlin Heidelberg.

Nguyen, H., & Masthoff, J. (2009). Designing empathic computers: The effect of multimodal empathic feedback using animated agent. In *Proceedings of the 4th international conference on persuasive technology* (pp. 7:1–7:9). ACM. https://doi.org/10.1145/1541948.1541958

Niewiadomski, R., Ochs, M., & Pelachaud, C. (2008). Expressions of empathy in ecas. In H. Prendinger, J. Lester, & M. Ishizuka (Eds.), *Intelligent virtual agents* (pp. 37–44). Springer Berlin Heidelberg.

Niewiadomski, R., Demeure, V., & Pelachaud, C. (2010). Warmth, competence, believability and virtual agents. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (Eds.), *Intelligent virtual agents* (pp. 272–285). Springer.

Niewiadomski, R., Bevacqua, E., Mancini, M., & Pelachaud, C. (2009, May). Greta: An interactive expressive eca system. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2* (pp. 1399–1400). International Foundation for Autonomous Agents and Multiagent Systems.

Nijholt, A. (2004, August). Where computers disappear, virtual humans appear. *Computers & Graphics*, *28*(4), 467–476. https://doi.org/10.1016/j.cag.2004.04.002

Noma, T., Zhao, L., & Badler, N. I. (2000, July). Design of a virtual human presenter. *IEEE Computer Graphics and Applications*, *20*(4), 79–85. https://doi.org/10.1109/38.851755

Novack, M., & Goldin-Meadow, S. (2015, September). Learning from gesture: How our hands change our minds. *Educational Psychology Review*, *27*(3), 405–412. https://doi.org/10.1007/s10648-015-9325-3

Nowak, K. L., & Biocca, F. (2003, October). The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, *12*(5), 481–494. https://doi.org/10.1162/105474603322761289

Ochs, M., Libermann, N., Boidin, A., & Chaminade, T. (2017). Do you speak to a human or a virtual agent? Automatic analysis of user's social cues during mediated communication. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (pp. 197–205). ACM. https://doi.org/10.1145/3136755.3136807

Ochs, M., Pelachaud, C., & Mckeown, G. (2017, January). A user perception–based approach to create smiling embodied conversational agents. *ACM Transactions on Interactive Intelligent Systems*, *7*(1), 4: 1–4:33. https://doi.org/10.1145/2925993

Parise, S., Kiesler, S., Sproull, L., & Waters, K. (1999). Cooperating with life-like interface agents. *Computers in Human Behavior*, *15*(2), 123–142. https://doi.org/10.1016/S0747-5632(98)00035-1

Patterson, M. L., & Sechrest, L. B. (1970). Interpersonal distance and impression formation. *Journal of Personality*, *38*(2), 161–166. https://doi.org/10.1111/j.1467-6494.1970.tb00001.x

Pejsa, T., Gleicher, M., & Mutlu, B. (2017). Who, me? How virtual agents can shape conversational footing in virtual reality. In J. Beskow, C. Peters, G. Castellano, C. O'Sullivan, H. Leite, & S. Kopp (Eds.), *Intelligent virtual agents* (pp. 347–359). Springer International Publishing.

Pejsa, T., Andrist, S., Gleicher, M., & Mutlu, B. (2015, March). Gaze and attention management for embodied conversational agents. *ACM Transactions on Interactive Intelligent Systems*, *5*(1), 3: 1–3:34. https://doi.org/10.1145/2724731

Pelachaud, C. (2005a). Multimodal expressive embodied conversational agents. In *Proceedings of the 13th Annual ACM International Conference on Multimedia* (pp. 683–689). ACM. https://doi.org/10.1145/1101149.1101301

Pelachaud, C. (2005b). Multimodal expressive embodied conversational agents. *Proceedings of the 13th Annual ACM International Conference on Multimedia* (pp. 683–689). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/1101149.1101301

Pelachaud, C. (2009a, December). Modelling multimodal expression of emotion in a virtual agent. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3539–3548. https://doi.org/10.1098/rstb.2009.0186

Pelachaud, C. (2009b, July). Studies on gesture expressivity for a virtual agent. *Speech Communication*, *51*(7), 630–639. https://doi.org/10.1016/j.specom.2008.04.009

Pelachaud, C. (2017). Greta: A conversing socio-emotional agent. *Proceedings of the 1st ACM Sigchi International Workshop on Investigating Social Interactions with Artificial Agents* (pp. 9–10). ACM. https://doi.org/10.1145/3139491.3139902

Pelachaud, C., Carofiglio, V., De Carolis, B., De Rosis, F., & Poggi, I. (2002). Embodied contextual agent in information delivering application. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 2* (pp. 758–765). ACM. https://doi.org/10.1145/544862.544921

Pelachaud, C., & Poggi, I. (2002, December). Subtleties of facial expressions in embodied agents. *The Journal of Visualization and Computer Animation*, *13*(5), 301–312. https://doi.org/10.1002/vis.299

Poggi, I., Pelachaud, C., De Rosis, F., Carofiglio, V., & De Carolis, V. (2005). Greta. a believable embodied conversational agent. In O. Stock & M. Zancanaro (Eds.), *Multi- modal intelligent information presentation* (pp. 3–25). Springer Netherlands. https://doi.org/10.1007/1-4020-3051-7%5F1

Poggi, I., Pelachaud, C., & De Rosis, F. (2000, November). Eye communication in a conversational 3d synthetic agent. *AI Communications*, *13*(3), 169–181.

Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., McCullough, K. E., & Ansari, R. (2002, September). Multimodal human discourse: Gesture and speech. *ACM Transactions on Computer-Human Interaction*, *9*(3), 171–193. https://doi.org/10.1145/568513.568514

Ravenet, B., Ochs, M., & Pelachaud, C. (2013). From a user-created corpus of virtual agent's non-verbal behavior to a computational model of interpersonal attitudes. In R. Aylett, B. Krenn, C. Pelachaud, & H. Shimodaira (Eds.), *Intelligent virtual agents* (pp. 263–274). Springer Berlin Heidelberg.

Ravenet, B., Clavel, C., & Pelachaud, C. (2018). Automatic nonverbal behavior generation from image schemas. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems* (pp. 1667–1674). International Foundation for Autonomous Agents and Multiagent Systems. http://dl.acm.org/citation.cfm?id=3237383.3237947

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge university press.

Rehm, M., & André, E. (2008). From annotated multimodal corpora to simulated human-like behaviors. In I. Wachsmuth & G. Knoblich (Eds.), *Modeling communication with robots and virtual humans* (pp. 1–17). Springer Berlin Heidelberg.

Rehm, M., André, E., Bee, N., Endrass, B., Wer, M., Nakano, Y., … Huang, H. (2007). The cube-g approach–coaching culture-specific nonverbal behavior by virtual agents. *Organizing and Learning through Gaming and Simulation: Proceedings of Isaga* (pp. 313).

Rickel, J., & Johnson, W. L. (1999, May). Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied Artificial Intelligence*, *13*(4–5), 343–382. https://doi.org/10.1080/088395199117315

Rickel, J., & Johnson, W. L. (2000). Task-oriented collaboration with embodied agents in virtual worlds. In *Embodied Conversational Agents* (pp. 29). MIT Press.

Rosis, F. D., Pelachaud, C., Poggi, I., Carofiglio, V., & Carolis, B. D. (2003, July). From greta's mind to her face: Modelling the dynamics of affective states in a conversational embodied agent. *International Journal of Human-Computer Studies*, *59*(1), 81–118. https://doi.org/10.1016/S1071-5819(03)00020-X

Salem, B., & Earle, N. (2000). Designing a non-verbal language for expressive avatars. In *Proceedings of the Third International Conference on Collaborative Virtual Environments* (pp. 93–101). ACM. https://doi.org/10.1145/351006.351019

Scherer, S., Marsella, S., Stratou, G., Xu, Y., Morbini, F., Egan, A., … Morency, L.-P. (2012). Perception markup language: Towards a standardized representation of perceived nonverbal behaviors. In Y. Nakano, M. Neff, A. Paiva, & M. Walker (Eds.), *Intelligent virtual agents* (pp. 455–463). Springer Berlin Heidelberg.

Schröder, M. (2010, January). The semaine api: Towards a standards-based framework for building emotion-oriented systems. *Advances in Human-Computer Interaction*, *2010*(2), 1. https://doi.org/10.1155/2010/319406

Schröder, M., Baggia, P., Burkhardt, F., Pelachaud, C., Peter, C., & Zovato, E. (2011). Emotionml – An upcoming standard for representing emotions and related states. In S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin (Eds.), *Affective computing and intelligent interaction* (pp. 316–325). Springer.

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. Wiley. (Google-Books-ID: Ze63AAAAIAAJ).

Sproull, L., Subramani, M., Kiesler, S., Walker, J. H., & Waters, K. (1996, June). When the interface is a face. *Human–Computer Interaction*, *11*(2), 97–124. https://doi.org/10.1207/s15327051hci1102_1

Stevens, C. J., Pinchbeck, B., Lewis, T., Luerssen, M., Pfitzner, D., Powers, D. M. W., Abrahamyan, A., Leung, Y., & Gibert, G. (2016, June). Mimicry and expressiveness of an eca in human-agent interaction: Familiarity breeds content! *Computational Cognitive Science*, *2*(1), 1. https://doi.org/10.1186/s40469-016-0008-2

Terven, J. R., Raducanu, B., Meza-de Luna, M. E., & Salas, J. (2016, January). Head-gestures mirroring detection in dyadic social interactions with computer vision-based wearable devices. *Neurocomputing*, *175*(2), 866–876. https://doi.org/10.1016/j.neucom.2015.05.131

Theune, M. (2001, December). Angelica: Choice of output modality in an embodied agent. In *International Workshop on Information Presentation and Natural Multimodal Dialogue (IPNMD- 2001)* (pp. 89–93). ITC-IRST.

Thórisson, K. R. (1997). Gandalf: An embodied humanoid capable of real-time multimodal dialogue with people. *Agents*, 536–537.

Tickle-Degnen, L., & Rosenthal, R. (1990, October). The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, *1*(4), 285–293. https://doi.org/10.1207/s15327965pli0104_1

Traum, D., Marsella, S. C., Gratch, J., Lee, J., & Hartholt, A. (2008). Multi-party, multi-issue, multi-strategy negotiation for multimodal virtual agents. In H. Prendinger, J. Lester, & M. Ishizuka (Eds.), *Intelligent virtual agents* (pp. 117–130). Springer Berlin Heidelberg.

Van Baaren, R. B., Holland, R. W., Kawakami, K., & Van Knippenberg, A. (2004, January). Mimicry and prosocial behavior. *Psychological Science*, *15*(1), 71–74. https://doi.org/10.1111/j.0963-7214.2004.01501012.x

Van Swol, L. M. (2003, August). The effects of nonverbal mirroring on perceived persuasiveness, agreement with an imitator, and reciprocity in a group discussion. *Communication Research*, *30*(4), 461–480. https://doi.org/10.1177/0093650203253318

Verberne, F. M. F., Ham, J., Ponnada, A., & Midden, C. J. H. (2013). Trusting digital chameleons: The effect of mimicry by a virtual social agent on user trust. In S. Berkovsky & J. Freyne (Eds.), *Persuasive technology* (pp. 234–245). Springer Berlin Heidelberg.

Wagner, D., Billinghurst, M., & Schmalstieg, D. (2006). How real should virtual characters be? In *Proceedings of the 2006 ACM Sigchi International Conference on Advances in Computer Entertainment Technology*. ACM. https://doi.org/10.1145/1178823.1178891

Wagner, H. L., MacDonald, C. J., & Manstead, A. S. (1986, September). Communication of individual emotions by spontaneous facial expressions. *Journal of Personality and Social Psychology*, *50*(4), 737. https://doi.org/10.1037/0022-3514.50.4.737

Walker, J. H., Sproull, L., & Subramani, R. (1994, April). Using a human face in an interface. In *Proceedings of the Sigchi Conference on Human Factors in Computing Systems* (pp. 85–91). Association for Computing Machinery. https://doi.org/10.1145/191666.191708

Wallbott, H. G. (1998). Bodily expression of emotion. *European Journal of Social Psychology*, *28*(6), 879–896. https://doi.org/10.1002/(SICI)1099-0992(1998110)28:6<879::AID-EJSP901>3.0.CO;2-W

Wang, C., Biancardi, B., Mancini, M., Cafaro, A., Pelachaud, C., Pun, T., & Chanel, G. (2020). Impression detection and management using an embodied conversational agent. In M. Kurosu (Ed.), *Human-computer interaction. multimodal and natural interaction* (pp. 260–278). Springer International Publishing.

Wang, I., Narayana, P., Smith, J., Draper, B., Beveridge, R., & Ruiz, J. (2018). Easel: Easy automatic segmentation event labeler. In *23rd*

International Conference on Intelligent User Interfaces (pp. 595–599). ACM. https://doi.org/10.1145/3172944.3173003

Wang, N., & Gratch, J. (2009, July). Can virtual human build rapport and promote learning? In *Proceedings of the 2009 Conference on Artificial Intelligence In Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling* (pp. 737–739). IOS Press.

Willis, F. N. (1966, June). Initial speaking distance as a function of the speakers' relationship. *Psychonomic Science*, *5*(6), 221–222. https://doi.org/10.3758/BF03328362

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). Elan: A professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 1556–1559). http://tla.mpi.nl/tools/tla-tools/elan/

Young, R. F., & Lee, J. (2004, August). Identifying units in interaction: Reactive tokens in korean and english conversations. *Journal of Sociolinguistics*, *8*(3), 380–407. https://doi.org/10.1111/j.1467-9841.2004.00266.x

About the Authors

Isaac Wang is a Ph.D. student studying Human-Centered Computing at the University of Florida. He earned his B.S. in Computer Science from Seattle Pacific University in 2015. His research focuses on gesture interactions with an emphasis on nonverbal communication in intelligent virtual agents.

Jaime Ruiz is an Associate Professor in the Department of Computer & Information Science & Engineering at the University of Florida. He received his Ph.D. in Computer Science from the University of Waterloo. His research is in the field of HCI, focusing on multimodal and natural user interfaces.