The Multimodal Nature of Embodied Conversational Agents

Max Louwerse (mlouwerse@memphis.edu)
Nick Benesh (nbenesh@memphis.edu)
Shinobu Watanabe (swatanab@memphis.edu)
Bin Zhang (bzhang@memphis.edu)
Patrick Jeuniaux (pjeuniau@memphis.edu)
Divya Vargheese (dvarghes@memphis.edu)

Department of Psychology / Institute for Intelligent Systems
University of Memphis
Memphis, TN 38152 USA

Abstract

Embodied conversational agents (ECA's) have become ubiquitous in human-computer interaction applications. Implementing humanlike multimodal behavior in these agents is difficult, because so little is known about the alignment of facial expression, eye gaze, gesture, speech and dialogue act. The current study used the data from an extensive study of human face-to-face multimodal communication for the development of a multimodal ECA, and tested to what extent multimodal behavior influenced the human-computer interaction. Results from a persona assessment questionnaire showed the presence of facial expressions, gesture and intonation had a positive effect on five assessment scales. Eye tracking results showed facial expressions played a primarily pragmatic role, whereas intonation played a primarily semantic role. Gestures played a pragmatic or semantic role, dependent on their level of specificity. These findings shed light on multimodal behavior within and between human and digital dialogue partners.

Keywords: embodied conversational agents, multimodal communication, avatars.

Introduction

Embodied conversational agents (ECA's) are animated characters that emulate human multimodal communication. Such communication involves both linguistic (e.g., speech intonation, discourse structure) and paralinguistic (e.g., facial expression, hand gestures, eye gaze) signals. There has been a baby boom of these agents both in the virtual world as well as in the literature. Offspring are produced in departments of psychology, artificial intelligence, computer science and education: human-like (Graesser, et al., 2004) and cartoon-like (Cassell et al., 2008); with anticipated careers in the military (Johnson et al., 2004), education (Graesser et al., 2004; McNamara, Levinstein, & Boonthum, 2004) or speech pathology (Massaro, 2006). In all these domains, the role of the agent is to improve the communication of a given message.

The genealogy of these agents goes back many centuries. One of the first proposals for embodied interfaces came from Heron of Alexandria's (62 AD) who described an 'automatic' puppet-theatre operated by weights. In the 18th century Friedrich von Knaus developed the first talking heads, while C. G. Kratzenstein synthesized vowel sounds using a set of acoustic resonators and vibrating reeds, and Von Kemplen developed the first speaking machine that produced sound combinations. In the first part of the 20th century Jacques Vaucanson developed mechanical animated objects like a flute-playing boy and a duck that could flap its wings, eat, and digest grain.

Obviously, lots of progress has been made, with today's ECA's being far more human-like than the older systems. At the same time, today's advanced embodied interfaces, like their predecessors, have a limited use of the multimodal aspects of communication. Even though some of today's systems have excellent speech interfaces (Pellom, Ward & Pradhan, 2000), conversational skills (Graesser et al., 2004), gestural movements (Cassell, Kopp, Tepper, Ferriman, & Striegnitz, 2007), or mouth movements (Massaro, 2006), they typically excel on just one aspect of multimodal communication. And even when human-like facial expressions and gestures are integrated in ECA's, they are carefully guided by literature but otherwise intuitive (Baylor & Kim, 2005), or come from actors acting out different modalities which are then transmitted to the agent, for instance by using body suits. The reason a full implementation of linguistic and paralinguistic channels of communication naturally used by humans has not been realized so far is that relatively little is known about how these channels combine within and across speakers in human-human communication. Although evidence has been collected on the alignment of pairs of modalities (e.g. Gullberg & Homqvist, 2006; Thompson & Massaro, 1996), few studies have investigated the associations between more than two modalities at a time.

In addition, it is questionable whether the development of, and research in, human-like ECA's is valuable in the first place. After all, the argument can be made that humans project human characteristics on objects that do not even slightly resemble humans (Reeves & Nass, 2003). Moreover, aiming for humanlike ECA's increases the chances of entering the uncanny valley (Mori, 2005), with users liking the humanlike avatar less than cartoonish avatars. At the same time, there is evidence that humanlike characteristics like stereotypes are applied to humanlike but not to cartoonlike agents (Louwerse, Graesser, Lu, & Mitchell, 2005).

Beyond the applied goal of improving the communication of a message to human users, ECA's can be employed for testing scientific hypotheses.

In a nutshell, it is important to uncover how multimodal channels are aligned in humans, and what the effect of alignment has on participants. That effect can ideally be tested using ECA's since they allow for careful manipulation of the linguistic and paralinguistic channels.

The aim of the current study is to use the data from a human-human multimodal communication experiment, implement these human facial movements, gestures and intonation in an ECA, and test the presence of each of the multimodal channels in a persona assessment, as well as in participants' attention to the agent.

Human-human communication

When language users communicate, they are involved in a rich complex of activities, involving discourse acts associated to the appropriate intonation and accompanied by facial expressions, hand gestures, and eye gaze. In a recent project on multimodal communication in humans and agents (Guhe & Bard, 2008; Louwerse et al., 2007) we collected 34 hours of multimodal dialogues from 64 students from the University of Memphis. Facial expressions and gestures were recorded by five camcorders, eye gaze was recorded by a remote eye tracker, and speech from both participants was recorded on separate audio channels.

To control base conditions, genre, topic, and goals of unscripted dialogs, we used the Map Task scenario (Anderson, et al., 1991). An Instruction Giver (IG) coached the Instruction Follower (IF) through a route on the map. By way of instructions, participants were told that they and their interlocutors had maps of the same location but drawn by different explorers and so potentially different in detail. They were not told where or how the maps differed, in order to increase the likelihood of observing diverse linguistic and paralinguistic signals.

Participants were seated in front of each other but were separated by a divider to ensure that they focused on the monitor. They communicated through microphones and headphones, and could see the upper torso of their dialogue partner and the map on a computer monitor in front of them through a webcam. This computer-mediated session, using webcams, was necessary for eye tracking calibration, as well as to reduce torso movement. The IG was presented with a colored map with a route (see Louwerse, 2007) and was asked to communicate the route to the IF as accurately as possible. The IF's task was to accurately draw the path on the screen using the mouse.

All dialogues were transcribed and each utterance was classified in one of 12 dialogue acts that are typically used for Map Task coding (Carletta et al., 1997; Louwerse & Crossley, 2006). Facial expressions were coded in a subset of the Action Units (Ekman, Friesen, Wallace, & Hager, 2002) and gestures were classified using McNeil's (1992) taxonomy.

Conventional statistical techniques like correlations and classical regression models are unsuccessful in determining the alignment of these communicative channels, because their use would assume that two variables are either fully synchronized on a time line or not at all. Moreover, the non-independence of observations would undermine the analysis based on these statistics. Instead, cross-recurrence analyses are useful because they can reveal the temporal dynamics of a data set and are meant to be used to model non-independent observations. Cross-recurrence plots quantify the recurrences of values in two times series. This nonlinear data analysis allows for comparisons between communicative channels as they unfold over time. This technique has been used successfully in illustrating the coupling of eye movements in dialog (Richardson, Dale, & Kirkham, 2007).

All modalities were at least polled at 250-millisecond intervals and a cross-recurrence analysis was run on this data. In addition, to identify whether the cross-recurrence pattern significantly differed from the baseline, a shuffled-time series baseline was computed. Only those multimodal channels were considered to be aligned if at least five points in the time series of the cross-recurrence analysis yielded a significant difference with the baseline as measured by a paired-sample t-test. Table 1 presents on overview of the alignment of the multimodal channels facial expressions, gestures and eye gaze to the dialog acts and should be interpreted as follows. For instance, when IGs use an Acknowledgment dialog act, they will keep their eyes on the map, and not on the IF. On the other hand, when IGs use an Explain dialogue act they look at the IF.

Implementation

Because of the applications embodied conversational agents are used in (e.g., intelligent tutoring systems), the ECA was developed to play the role of the IG, while the human user would play the role of the IF. This meant that the agent was developed to communicate the path on the map to the IF. The agent program was developed using Visual C# and Visual C++ in Visual Studio 2005. It consisted of two main components, the interface program and a speech recognition system, which communicated through a TCP socket.

The interface program had a full screen dialog window divided into two halves. On the left half, a *Haptek* avatar was situated. On the right half, the IF map. A dialog manager decided what the avatar said and how the avatar behaved. The dialog manager simulated a state machine. Within each map location, the dialog manager created states. Each state served a dialogue function: 1) confirm current location, 2) give instruction, or 3) back up to previous location. For every state change the dialog manager first took input from the *LumenVox* Speech Engine, processed this information and produced a response using *Speechify* speech synthesis and *Haptek* facial expressions, eye gaze and gestures.

	eye on IF	eye on map	blinking	squint	eye brows up	asymm. brows	eye brow down	frown	lip tightener	mouth open	pucker	biting lip	smile	laugh	nodding	shaking	stroke face	chinrest	deictic gesture	iconic gesture	route gesture	multi beat	single beat
acknowledgment	-	+	+	-			-			-		+	+	+	+	-	+		-	-	-	-	-
align	+																						
check					-								-										
clarify												-			+	+	-		+				
explain	+		+	+	+			+	-	-	-	-	+		+	+	-		+	+		+	+
instruct	-	+	+		+	+			-	-	-	-	-	-	+	+	-	-	+	+	+	+	+
query-w			+	+				+	-	-		-								-			
query-yn	+			+	+		+		-	-	-	-	-	-	-		-		+				+
ready	-			-		-	-		-			-			-	-			-	-			
reply-n					+							-				+	+						
reply-w					+					-	-	-	-		+				+		-		-
reply-y	-		+		+	-			-	-					+				-	-		-	
	eyes				eye brows				mouth						head		hands						

Table 2. Overview of cross-recurrence patterns between dialogue acts and other modalities. A positive cross-recurrence (+) indicates a higher, and a negative cross-recurrence (-) indicates a lower frequency of events, compared to the baseline.

To create the facial movements for the *Haptek* agent (Figure 1), the Action Units (AUs) linked to the selected facial expressions were taken as templates. Activation of the AU was based on IG cross-recurrence behavior. Facial movements and gesture movements worked on a pre-set muscle and a joint point system. This allowed for natural multimodal behavior to be implemented. The intensity of behavior was modified when considered too expressive (or unnatural) based on trial and error testing to achieve desired effect.

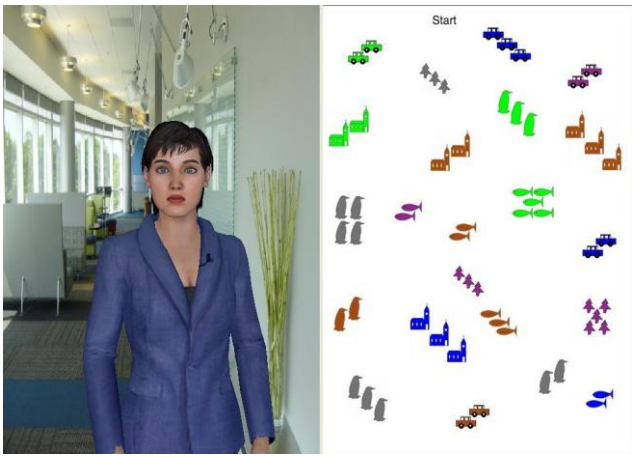

Figure 1. Interface with *Haptek* agent and IG map.

Experiment

An experiment tested whether the modalities implemented in the agent had a positive effect on the perceived usefulness of the agent and the performance at the task. In order to test the impact of the naturalness of the agent conditions were created whereby facial expressions were (or were not) activated, gestures were (or were not) activated, and intonation was (or was not) activated. The intonation condition used the Speechify intonation or removed any intonation that the Speechify synthesized speech uses. This resulted in 2 (face) x 2 (gesture) x 2 (intonation) = 8 within-subject conditions. In addition, two intonation specific conditions were added without face or gesture movements, but with enhanced intonation that either matched or mismatched Steedman's (2000) theory of contrast (correct vs. incorrect stress).

In an eye tracking experiment using rather static agents Louwerse et al. (in press) found ECA's to attract attention to the nose bridge. Their findings were very similar to Gullberg and Holmqvist (2006) who reported eye tracking evidence that the face of the dialogue partner dominates as a target of visual attention, whereby fixations would primarily center on the nose bridge of the speaker's face capturing the eyes and mouth of the speaker simultaneously. This suggests that the face fulfills a pragmatic role. We therefore predicted that the same pragmatic effect would emerge in the face condition.

In an eye tracking experiment on pointing gestures and linguistic expressions Louwerse and Bangerter (2005) found

gestures fulfilling very much a semantic function. When information from linguistic expressions did not suffice, attention moved to gestures. However, when linguistic expressions were sufficient, gestures did not receive the same amount of attention. We therefore predicted that the same semantic effect would emerge in the gestures and intonation condition.

Participants
Twenty-four students at the University of Memphis interacted with the ECA and received course credit for their participation.

Materials
Fourteen maps were used, 10 experimental maps and 4 filler maps. The order of the 8 (2 x 2 x 2) multimodal maps was fixed but the order of the conditions was counterbalanced. The order of the two intonation specific maps was counterbalanced. In the multimodal maps, all modalities (face, gesture, and intonation) were varied per condition. In the intonation specific maps, only stress was varied between correct and incorrect. Maps were of equivalent difficulty, and similar to those in the human-human experiment discussed before. As in the human-human experiments, there were slight differences between the IG and IF maps to elicit conversation.

Apparatus
Participants' communication was recorded by camcorders and a speech recorder, similar to the set up in the human-human experiment. We will focus here on eye gaze only, recorded for the IF using an SMI iView RED remote eye tracker with a sampling frequency of 60Hz.

Procedure
For all 14 maps, participants were seated in front of the computer presenting the ECA. They communicated through a microphone and headphones with the ECA. In between maps, the eye tracker was recalibrated to ensure precision.

After completing each map, participants filled out a questionnaire based upon Ryu and Baylor's (2005) Agent Persona Instrument to evaluate the ECA in the relevant condition. This instrument is the result of factor analyses on data from a number of human-agent interaction studies and consists of questions related to four categories (*facilitation of learning, credibility, human-likeness and* e*ngagement*). In addition, we added questions related to the extent participants liked the quality of the interaction (e.g., I liked the agent's voice; I liked the agent's appearance). All questions were answered on a 1-6 scale, 1 being *totally disagree* and 6 being *totally agree*.

Results

Questionnaire
Internal consistency of the questionnaire as measured by Cronbach's α was computed on all 24 participants. Overall reliability was .86. High internal consistency was found for all five categories, facilitation of learning (α = .80), credibility (α = .92), human-likeness (α = .82), engagement (α = .83), and quality (α = .90).

We conducted 2 (presence/absence of facial expressions) x 2 (presence/absence of gestures) x 2 (presence/absence of intonation) mixed-model analysis on the participants ratings with participants and items as random factors (Baayen, Davidson, & Bates, 2008). The model was fitted using the restricted maximum likelihood estimation (REML) with a Kenward-Rogers adjustment for degrees of freedom.

The presence of facial expressions had a positive effect on answers in all five categories, the presence of gesture on answers in all five categories except credibility. Intonation positively affected all five categories except the categories humanlike and credibility. Results are presented in Table 3. None of the enhanced correct or incorrect stress conditions yielded significant differences. These findings first and foremost suggest participants value multimodal behavior in ECA's, whereby the role of facial expressions is most important. Gestures and intonation play a slightly lesser role particularly when it comes to assessment of credibility (gesture and intonation) and human-likeness (intonation). These findings might support the specifically pragmatic role for facial expressions and the specifically semantic role for gestures and intonation. However, testing semantic factors requires a measurement other than a persona assessment. We therefore looked at the role of eye fixations in the interaction.

Eye gaze
Areas of interest (AOI) were defined as areas on the face of the ECA, the start and end locations on the map, and items important for disambiguation of location based on shape or color. Total fixation time on areas of interest on the ECA and the map were computed. Outliers were defined as 3 SD above the mean within a condition, subjects and area of interest, and were removed from the analysis. This affected less than 3% of the data.

As before, a mixed-effects model was used with the total fixation time as the dependent variable and with participants and items as random factors and presence and absence of face, gesture and intonation as fixed factors.

The presence of facial expressions increased the fixation time on the face of the agent and more specifically on the nose bridge of the agent. This finding is in line with eye tracking studies in human-human communicative settings discussed earlier, and confirms the hypothesis that facial expressions play a pragmatic role in interactions.

The increased fixations on the ECA's face cannot be explained by the fact that motion attracted the attention, as the presence of gestures also directed attention to the face of the ECA, even in the absence of facial expressions.

	Face		gesture		Intonation	
	absence	presence	Absence	presence	absence	Presence
facilitation learning	3.01 (1.59)	3.91 (1.54)**	3.11 (1.62)	3.78 (1.56)**	3.17 (1.61)	3.76 (1.60)**
credibility	3.17 (1.79)	3.81 (1.58)**	3.26 (1.80)	3.69 (1.61)	3.33 (1.76)	3.69 (1.64)
human-likeness	3.31 (1.54)	3.79 (1.44)**	3.35 (1.53)	3.73 (1.47)**	3.40 (1.53)	3.70 (1.48)
engagement	3.03 (1.57)	3.96 (1.48)**	3.14 (1.60)	3.81 (1.51)**	3.18 (1.58)	3.80 (1.54)**
quality	3.10 (1.59)	3.63 (1.56)**	3.15 (1.59)	3.56 (1.58)**	3.22 (1.59)	3.55 (1.59)*

*Table 3. Means and standard deviations of ratings in persona assessment questionnaire. ** p < .01, * p < .05*

AOI	face		gesture		intonation	
	absence	presence	absence	presence	absence	presence
start	199.66 (186.66)	185.67 (176.79)	181.99 (180.23)	212.93 (185.75)*	180.88 (157.60)	187.66 (176.64)
end	147.47 (128.23)	105.31 (102.30)**	136.23 (126.43)	123.70 (111.22)	127.83 (122.02)	138.56 (118.88)
face	1452.71 (936.55)	1824.66 (1352.61)**	1519.46 (1176.12)	1729.30 (1068.15)*	1636.32 (1294.45)	1570.76 (1034.81)
nose	269.48 (264.66)	379.88 (472.63)**	343.79 (427.79)	280.96 (278.30)	311.68 (430.40)	293.38 (325.63)
eyes	358.71 (362.11)	518.03 (721.24)*	439.91 (633.26)	407.04 (409.00)	447.18 (684.30)	410.05 (437.31)
color	467.09 (465.27)	560.1 (528.78)	460.87 (445.91)	574.68 (554.04)	602.50 (599.60)	478.92 (408.87)*
shape	614.77 (563.49)	502.84 (456.43)*	587.00 (545.36)	543.39 (493.93)	544.39 (454.72)	592.75 (626.11)

*Table 3. Means and standard deviations of total fixation times on areas of interest (AOI). ** p < .01, * p < .05*

This finding does not confirm the hypothesis that gestures play a semantic role, but perhaps the semantic content was too general. The finding, however, is in line with the human-human communication literature, that shows that addressees often do not attend to gestures but instead fixate on the face of the dialogue partner (Gullberg & Holmqvist, 2006).

Facial expressions and gestures also played a role at the start and the end of the experiment. Fixations on the opening landmark on a map received more fixation time when gestures were present, while closing landmarks received more fixation time when facial expressions were present. This might suggest that at the start of the map navigation addressees need hand gestures to get oriented (semantic factors), whereas at the end of a map these gestures are not needed but eye contact to close the dialogue is (pragmatic factors).

The area of interest shape referred to one specific group of three landmarks on each map for which shape was a disambiguating factor. For instance, in a situation involving two blue fish, two red cars and two red trees in each other's vicinity, the use of the referential expression *two red trees* the disambiguating word referred to the shape. Similarly, the area of interest color referred to one specific group of three landmarks on each map for which color was the disambiguating factor. Because linguistically color always preceded shape, color provided slightly more ambiguity. In the absence of any intonation, fixation times indeed increased on these color landmarks, when the participant had to compare the correct landmark of two similar ones. This confirms the semantic role for intonation. No differences in fixations to shape were found.

The role of intonation in disambiguation was also found when incorrect and correct stress was compared. In the case of incorrect stress, fixation time was three times higher on the color landmarks than when correct stress was given (M = 482.14, SD = 358.89 vs. M = 190.56, SD = 213.08, $F(1,23) = 12.29$, $p < .01$). Recall that in the conditions of incorrect and correct stress no facial expressions or gestures were present. Nevertheless, a difference approaching significance was found between intonation conditions in fixation time on the area of the gestures, with twice as much fixation time on the gesture area in the incorrect stress condition than in the correct stress condition (M = 605.88, SD = 681.83 vs. M = 319.68, SD = 366.96, $F(1, 46) = 3.69$, $p = .06$), as if incorrect stress made the need for gestural cues larger (Louwerse & Bangerter, 2005). These findings confirm the semantic role of intonation, and that of specific gestural movements, in communication.

Conclusion

The current study investigated the multimodal behavior in ECA's. Two questions played a central role, the first being how multimodal behavior can be implemented in ECA's if so little information is available on the alignment of communicative channels in humans; the second what the effect of humanlike multimodal behavior is on interactions with ECA's. Using a large multimodal corpus of face-to-face conversations, we were able to implement natural humanlike multimodal behavior in ECA's. That this implementation was perceived as being efficacious was confirmed in the assessment of the persona. Moreover, interactions with the ECA showed that facial expressions, gestures and intonation all had a positive effect on the

communication, with some evidence that facial expressions played a pragmatic role, whereas intonation played a semantic role. Gestures had a pragmatic factor when they were general, a semantic factor when they were specific. These findings shed light on multimodal behavior within and between human and digital dialogue partners.

Acknowledgments

This research was supported by grant NSF-IIS-0416128. We thank Ellen Bard, Rick Dale, Markus Guhe, and Mark Steedman for their help in the experiment design and data analysis. We also express our thanks to *Haptek* and *LumenVox* for their support. The usual exculpations apply.

References

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., & Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech, 34*, 351-366.

Baayen, R.H., Davidson, D.J. and Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390-412.

Baylor, A. L. & Kim, Y. (2005). Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education, 15*, 95-115.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., & Anderson, A. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics, 23*, 13-31.

Cassell, J., Kopp, S., Tepper, P., Ferriman, K., & Striegnitz, K . (2007) Trading spaces: How humans and humanoids use speech and gesture to give directions. In T. Nishida (ed.) *Conversational Informatics* (pp. 133-160). New York: John Wiley & Sons.

Ekman, P., Friesen, Wallace V., & Hager, J.C. (2002). *Facial Action Coding System (FACS)*. CD-ROM.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: a Tutor with Dialogue in Natural Language. *Behavior Research Methods, Instruments, and Computers, 36*, 180–193.

Guhe, M. & Bard, E. G. (2008) Adapting the use of attributes to the task environment in joint action: Results and a model. *Proceedings of Londial – The 12th Workshop on the Semantics and Pragmatics of Dialogue*, 91–98.

Gullberg, M. & Holmqvist, K. (2006). What speakers do and what listeners look at. Visual attention to gestures in human interaction live and on video. *Pragmatics and Cognition, 14*, 53-82.

Johnson, W. L., Choi, S., Marsella, S., Mote, N., Narayanan, S., & Vilhjálmsson, H. (2004). Tactical language training system: Supporting the rapid acquisition of foreign language and cultural skills. *Proceedings of InSTIL/ICALL*.

Louwerse, M.M. & Bangerter, A. (2005). Focusing attention with deictic gestures and linguistic expressions. In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Cognitive Science Society* (pp. 1331-1336). Mahwah, NJ: Lawrence Erlbaum.

Louwerse, M. M. & Crossley, S. A. (2006). Dialog act classification using N-Gram algorithms. In G. Sutcliffe & R. Goebel (Eds.), *Proceedings of the International Florida Artificial Intelligence Research Society* (pp. 758-763). Menlo Park, California: AAAI Press.

Louwerse, M. M., Benesh, N., Hoque, M.E., Jeuniaux, P., Lewis, G., Wu, J., & Zirnstein, M. (2007). Multimodal communication in face-to-face conversations. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 1235-1240). Austin, TX: Cognitive Science Society.

Louwerse, M.M., Graesser, A.C., Lu, S., & Mitchell, H.H. (2005). Social cues in animated conversational agents. *Applied Cognitive Psychology, 19*, 1-12.

Louwerse, M.M., Graesser, A.C., McNamara, D.S. & Lu, S. (in press). Embodied conversational agents as conversational partners. *Applied Cognitive Psychology*.

Massaro, D.W. (2006). A computer-animated tutor for language learning: Research and applications. In P.E. Spencer & M. Marshark (Eds.), *Advances in the spoken language development of deaf and hard-of-hearing children* (pp. 212-243). New York, NY: Oxford University Press.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy trainer for active reading and thinking. *Behavioral Research Methods, Instruments, and Computers, 36*, 222-233.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago, IL/London, UK, University of Chicago Press.

Mori, M. (1970). The uncanny valley. *Energy, 7*, 33–35.

Pellom, B., Ward, W., & Pradhan, S. (2000). The CU communicator: An architecture for dialogue systems. *International Conference on Spoken Language Processing (ICSLP)*. Beijing, China.

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

Richardson, D.C., Dale, R. & Kirkham, N.Z. (2007). The art of conversation is coordination: Common ground and the coupling of eye movements during dialogue. *Psychological Science, 18*, 407-413.

Ryu, J. & Baylor, A. L. (2005). The psychometric structure of pedagogical agent persona. *Technology, Instruction, Cognition & Learning, 2*, 291-319.

Steedman, M. (2000). Information structure and the syntax-phonology interface. *Linguistic Inquiry, 34*, 649-689.

Thompson, L. & Massaro, D. (1986). Evaluation and integration of speech and pointing gestures during referential understanding. *Journal of Experimental Child Psychology 42*, 144- 168.