Whole Body Interaction for Child-Centered Multimodal Language Learning

Berto Gonzalez
berto@u.northwestern.edu

John Borland
j-borland@northwestern.edu

Kathleen Geraghty
k-geraghty@northwestern.edu

Center for Technology and Social Behavior
Northwestern University
Evanston, IL 60208

Abstract
Children engage with the world with their whole bodies, and we suggest here that during dialect learning, as during other learning activities, technology be capable of responding in whole body ways. As the child becomes more engaged in a shared-reality environment, the coordination of the whole-body behaviors between the VP and child should increase, thereby enhancing the experience. In this paper, we present our work on developing a virtual agent that embodies whole-body behaviors and a shared-reality environment that encourages children to use whole-body expression in the context of learning dialect, and science talk.

Categories and Subject Descriptors
H.5.1 [**Multimedia Information Systems**]: Artificial, augmented, and virtual realities

General Terms
Design, Human Factors

Keywords
Embodied Conversational Agent, Culture, Analysis and modeling of verbal and nonverbal interaction

1. INTRODUCTION
The achievement/opportunity gap between African American and Euro-American children is well known and persistent in the American educational system [1].We propose a technology to address this gap, in the domain of science learning by engaging African American English (AAE) speaking children in a bridge building and discussion task with a Virtual Peer (VP) that allows the children to explore the use of Mainstream American English (MAE), at the same time as they engage in the evidence-based science talk that classrooms prize. But, as Crowder (1993) shows, children's science talk also includes the use of the body [2]. We believe that a shared-reality environment with a VP can leverage learning by encouraging these whole-body behaviors.

ICMI-MLMI'09 Workshop on Child, Computer and Interaction
November 5, 2009, Cambridge, MA, USA

2. BACKGROUND
The innovation of the interface is that children actually engage in collaboration with the virtual peer, as they use instrumented Legos to build a Lego bridge that will meet in the middle, and that they switch roles from being peers to taking turns playing the teacher. The innovation of the design process is that we base our work on a deep analysis of a long period of data collection with the community of interest. In this particular paper we discuss a new way of processing the corpus of human-human data in order to implement behaviors for the virtual agent.

We first observed 40 African American and Euro-American children engaging in a collaborative bridge building exercise using Lego® Duplos. The children were instructed to build a bridge strong enough and wide enough to allow "people" (small figurines) with "bags of food" (weighted cloth bags) to cross a dangerous river. Once the bridge was finished, the children were told that they might want to practice their sharing time description of what they had done by alternating playing the teacher, asking or answering questions about the dyad's choices in designing and building the bridge. We transcribed and annotated the interactions, to create a child-child corpus. This corpus serves as the basis for the VP's verbal and non-verbal behavior, and we believe that this tight loop between the observation of real children and the design of virtual children will ensure our VP conveys culturally authentic peer-like behavior that elicits whole-body behavior on the part of a real child user.

The current workshop paper is a report of work in progress as we experiment with different methods of processing the corpus using machine learning techniques, and we are currently in the final stages of implementing a working shared reality system composed of a multimodal Virtual Peer, a Wizard of Oz (WoZ) interface driven by statistical models, and context sensors on the Lego pieces and table on which the children build.

2.1 Virtual Peer
The agent that plays the role of peer was designed to be ethnically-ambiguous [3] and to suggest the age of a 3rd-grade student. As we have shown, ethnic ambiguity is possible, and it allows us to manipulate the perceived ethnicity of the VP based solely on its verbal and non-verbal behaviors, rather than imposing our stereotypes of ethnicity [3]. The VP can produce (recorded human) speech in two dialects of English: Mainstream American English (MAE) and African American English (AAE). The VP code-switches (shifts its dialect), based on the context. For bridge building, the agent speaks in AAE. During the teacher-student classroom phase, the agent speaks MAE. The goal is to

elicit the use of "classroom English" in the context of science explanation. Because the VP embodies a culturally-authentic peer that uses head, body, and language in ways observed during real children's interaction, we find that children who play with the peer tend to also express their understanding of science concepts through a combination of linguistic skills and whole-body movements, as they do with their real peers.

2.2 Wizard of Oz

An ideal virtual peer system of this sort would be completely autonomous, and could serve as a stand-alone system in a classroom. However, current technologies in the fields of speech recognition, Natural Language Understanding, and Text-to-Speech cannot reliably understand or believably generate the language of AAE-speaking children. There are indications that adaptations can be made to help with children's speech recognition [4], [5] but these modifications are beyond the scope of our current effort. Our goal, therefore, is to create a VP that is semi-autonomous, where a human operator plays the role of speech recognition and natural language parsing through the use of a Wizard of Oz (WoZ) interface, with pre-recorded speech output.

Although we find it necessary to involve a human operator in the input process, we want to limit as much as possible the degree to which his/her subjective judgments about how children behave influence the VP's behavior. Therefore we integrate conversational planning through statistical models that take input from context sensors and from the WoZ interface, and generate the VP's system output in the form of Behavior Markup Language (BML) [6]. The first stage of the statistical model is a Markov Model (MM) generated off-line from the corpus of interactions between AAE-speaking children performing the same bridge building and teacher-student conversation. We annotated this corpus with utterance categories (e.g. Question, Exclamation, etc) and calculated transitional probabilities for each utterance category given Markov chains of preceding turns in the conversation. Given these chains as input, the MM selects the next utterance category based on the distribution found in the corpus. The calculations of these probability distributions are explained in greater detail in the next section. During an interaction with a child, the WoZ operator listens to what the real child says, and then quickly annotates the input with a predefined set of utterance categories by pressing the respective button on the WoZ. The WoZ then displays a small set of utterances that a child might have responded with in the same context, and the WoZ operator chooses one of these utterances as a response. Given that there is no propositional understanding, the fact that there is a set of utterances gives the operator the ability to pick an utterance more relevant to the previous utterance.

Once the utterance has been chosen, gestures and head movements appropriate to the utterance are looked up in an assignment table calculated off-line probabilistically selected by the system. Gestures are associated directly with individual utterances, since they are often specific to what is being said. For example, a waving gesture is associated with the "Hi, I'm Alex" utterance. A more general gesture, such as pointing, might be associated with several utterances according to where it appears in the corpus. The gestures are probabilistically weighted according to the corpus as well.

Eye gaze for an utterance is selected according to a statistical model based on utterance type and the context This more general method is supported by work such as [7], which links gaze behaviors to utterance types. This information is calculated directly from the corpus for eye gaze at the beginning, middle and end of Alex's turns as speaker. Most turns were too short for the motion planner to blend three animations together, so we implemented gaze targets at utterance beginning and end.

Just as important as the nonverbal accompaniments to speaking behavior are Alex's listener behaviors. Our future plans for the system include incorporating eye gaze detection, as has been done in previous systems, in order to generate appropriate responses when Alex is listening [8], [9]. However, as the system is still under development, our current strategy is for the WoZ operator to manually cue verbal and nonverbal feedback by pressing buttons on the WoZ.

2.3 Data Model

We are currently experimenting with choosing the virtual peer's next utterance by building off-line a Markov Model chain of utterances in the child-child corpus. We then look for similar chains of utterance types and predict Alex next utterance on the basis of the child-child corpus. Since we are concerned with generating the next system utterance after a turn boundary, the history contains utterances immediately before turn boundaries. The chains used to construct the model consist of such histories of utterances before turn boundaries followed by the first utterance after the last turn boundary for each chain. Using this information, we can construct probability distributions of what utterance types are likely to occur given every possible history chain.

In building such a system, given the exigencies of transcribing 15 minutes of child-child conversation along with non-verbal behavior, the issue of sparseness of data must be addressed. For example, a 10-15 minute child-child interaction generally includes about 300 utterances per child. We are using a history of length 3 to predict an utterance, so this means that the Markov chains used to construct the model are of length 4 (a history of 3 plus the first utterance to be emitted after the history). The chains are based on utterance type categories 15 (or more, depending on the task). So a 4^{th} order chain with 15 utterance types gives $15^4 = 50{,}625$ possibilities. Even if our data sets consisted of conversations that counted in the tens of thousands of utterances, there would still be a good chance that chains not present in our child-child corpus would occur during child-VP interaction. Therefore, we have to use a method that allows us to predict utterances for unseen conversation histories. To solve this problem we are trying out a Katz backoff model with Good Turing estimation, as described in [10]. We used the SRILM toolkit [11] to implement this model.

The Katz-Backoff model is a generative model that is popular in speech recognizers. This model can compensate for sparse data by considering longer event histories when they are represented in the data, and backing off to shorter histories when they are not. This method necessitates the proper distribution of probability mass, since we are considering event sequences that have not been seen and therefore are not accounted for in a direct probability distribution. Good-Turing estimation provides such a method, by smoothing the probability distribution in such a way that unseen events are assigned a nonzero probability.

2.4 Context Sensors

Some of the output utterances have tags that indicate that they are specific to the physical context. For example, "Use the blue block" requires that there be a blue block on the table. Because

the WoZ operator's attention will already be focused on quickly annotating the child's speech, and won't have time to scan the scene, we are currently in the process of equipping our system with sensors that can specify whether pre-defined contextual conditions hold true (e.g. "Is there an *X* colored block?" or "Is the bridge partially built?").

Patel, Bosley et al, (2006) demonstrate that situational awareness can improve believable social experiences with robots [12]. The contextual awareness of their robot was achieved using software controlled by a human. For our situated environment, we believe giving our VP contextual awareness that relates to bridge building (e.g. progress of the bridge, color of blocks, location of blocks on table, width vs. height of bridge, etc) is important in engaging a child participant. Our VP's awareness of the physical context will result from sensors on the blocks and table.

During the early stages of development, we used Infrared LED's and a Nintendo Wii Remote to track the movement of a Lego block moving across a table in the real world. As the block moved across the table, real-world coordinates were collected by the Wii Remote and reported to the VP System. The system could then convert these coordinates to virtual-world coordinates that the VP's gaze behavior (i.e. the direction the VP's head and eyes are pointed) could follow. Because of the amount of computation required to accurately and smoothly track a single object with gaze and head movements, we felt that this very dynamic behavior was actually limiting, and therefore did not integrate this sensor in our experiments.

Currently, we have been using object-tracking software and a camera pointed toward the table space where the bridge is built. It provides information about the objects in view such as size, position, movement, and proximity to one another. It "blobs" objects in close proximity to form a single object (i.e. when blocks that were once separate are connected, they become one object in the system). Using this data, we can make interpretations about the current context using predefined criteria. For example, if there have been no new blocks placed on the table for a while, the system can either assume the child is stuck and in need of suggestions or that the child has completed his/her bridge.

For the next iteration of context sensors, we are looking at using RFID tagging. Our hope is that RFID will be as informative as the object tracking software, but require less computation.

2.5 Tangible Interface

With sensors, the Lego blocks are objects of play, but also a tangible interface for interaction between the real and virtual peer. Barakonyi, et al (2005), extend Augmented Reality (AR) by providing virtual agents with contextual awareness. Cameras are used to generate information about objects in the context. For example, the virtual repairman assists the user in building a robot. The agent overlays a virtual piece in the AR at the position the real piece should be placed. By looking at the context, the virtual agents can provide feedback to the user [13]. Our system relies on multimodality in the form of nonverbal embodied output for the agent, but also bridge-building actions by the child and agent. Data on what stage in the bridge building process the child is at and how successful or unsuccessful the child has been in designing and building the bridge is analyzed by the system to generate response behaviors. For example, if the child has been unsuccessful at building a stable bridge, the VP's gaze may shift from its own blocks to the child's blocks, and it might suggest a strategy. Similarly, in Gebhard and Klesen (2005), the agent's discourse is affected by what the user has built [13, 14]. Likewise, sensors change the state of the VP's dialogue based on the current state of the child's bridge. For example, it would not be appropriate for the VP to comment on the integrity of the child's bridge until the child has a freestanding bridge. In this way, the child is influencing what sort of discourse ensues not just by what is spoken but also by his/her actions.

The blocks also serve as a mode for generating reactive behavior, or behaviors that occur between 0 - 1 seconds [15]. The sensors can detect quick movements that a real person would react to automatically, without thinking, such as if the bridge collapses in a heap during testing. The system will have a reactive layer [15] that can then respond to the quick change in context by having the agent look in the direction of the child's bridge.

The work with the Wii Remote software, mentioned earlier, led us to create important invisible objects or "gaze targets" in the virtual world that would map to real world objects. One such gaze target is the "child's table", which directs the VP's eyes and head toward a point in the virtual world that makes the VP appear to be looking at the table in the real world. Using these gaze targets, the data from the sensors, and the predefined criteria, the reactive behavior of responding to a bridge collapsing becomes possible in a virtual agent.

We are continuing to work on criteria for reactive behaviors that we can observe and determine to be important in our child-child corpus.

3. DISCUSSION

Interactions with a virtual agent displaying and reacting to whole body interactions are not new [16, op cit], and are enjoyed by participants, as well as giving them a greater sense of co-presence with the agent [17]. In the work we are pursuing here whole-body interaction includes language, nonverbal behavior, the actions of collaborating on building a bridge, and the responses of each agent (human and virtual) to each of these forms of embodied behavior on the part of the other.

In the next sections we describe observations from our initial pilot studies that support our work in the direction of action as well as nonverbal linguistic behavior in with VPs, especially in learning contexts.

3.1 Rigid vs. Whole Body

[17] found many participants had a rigid stance when first immersed in a "virtual reality". As participants realized the response capabilities of the system, their body movements relaxed. Similarly, children first introduced to the VP, stood rigidly in front of the screen, staring at the projected image. As children observe the VP interact with its environment (i.e. the virtual blocks), they are no longer solely looking at the VP, they look at what is happening in the VP world. Blocks are removed from a bucket and a bridge is being built. Children quickly relax and in fact act as they might with a close friend ("it's not that I'm bored," said one child to the VP, "I just don't know what to build next").

Children's body movements begin to synchronize to the VP's movements. For example, when the VP picked up a bag of "food" to test the bridge, one child asked the VP to wait and while the VP was holding the bag just above the table, the child hurriedly finished her bridge and then held her weighted bag above the bridge so that they could test their halves of the bridge at the same

time. The child recognized that the VP was waiting and responded by working more quickly.

We also found that children synchronized non-bridge-building aspects of their behavior. At one point in the interaction, the VP sings or hums a familiar song featured in well-known movies and covered by gospel artists (that we heard in our child-child corpus) and then fades out. Virtually all of the children pick up the singing and continue the next verse. Later in the interaction, those same children are likely to sing or hum another tune and do a small dance without it being initiated by the VP.

In the child-child corpus, embodied behaviors around the table did take a different form than when they were interacting with the VP, along one dimension. Some children stood atop their chairs to get better views, while others walked around the whole table to get access to the whole play surface. Because of the VP's projection screen, children are a little more restricted, but still move to the different sides of the table and utilize their whole body to express themselves, including clapping to celebrate a solid bridge that successfully held the weighted bags.

Interestingly, fewer embodied behaviors occur during the teacher-student task, when children tend to sit with a straight back and folded hands, as they would with a real teacher.

As discussed earlier, our VP generates non-verbal behaviors according to the probabilistic distribution of occurrence in our child-child corpus. Foster and Oberlander (2007) show that agents whose non-verbal behavior occurred on the basis of a distribution of actions that occurred in their corpus were more liked as compared to agents whose non-verbal behavior occurred as a function of selecting the non-verbal behavior with the highest rate of occurrence in their corpus [18]. These distribution-based behaviors make the VP appear less robotic in that participants cannot predict the non-verbal behavior of the VP.

Children become highly attuned to what our VP can and cannot do. Before we incorporated a model of eye-gaze that included gaze at the other child and at the actions of the child with the Lego blocks, and decided to build block sensors into the system, one child took a block and held it in the air in front of her and to the right. She stared directly at the VP and stated, "Look at this." She was testing the VP's reaction to see if the VP would look at the block or perhaps follow the child's eye movements. Another child exclaimed, "Did you see that […]?" after his bridge had fallen, and then looked at the VP. The VP was looking straight ahead, instead of at the fallen bridge. Both children appeared disappointed that their expectations of how the VP should act or react were violated, and no longer looked at the VP as they were playing, much like the results in [7]. This suggests that the shared reality encourages children to respond with "playing and talking in familiar ways" [19]. While the technology in VP systems is often limited and children will inevitably find the shortcomings of the system, VP systems needs more whole body behavior, including action as well as linguistic behavior. Children become more engaged as they discover the Virtual Peer can effectively embody behavior that is familiar to the child.

4. CONCLUSION

We believe that our reliance on a child-child corpus allows our VP to believably model peer-like culturally-specific language and nonverbal behaviors. By so doing, it situates the child user in an environment in which they are encouraged to use their whole body as a scaffold to learning, and to follow the VP's lead in switching dialects between the bridge building and discussion/ classroom tasks. For example, children were observed enacting physical forces with the use of their movements. One child said, "So it would not fall." On the word 'fall', she put her arms overhead and swung them down next to her sides while bending her knees, demonstrating her understanding of forces she observed through her own physical embodiment. Pilot data also shows that AAE-speaking children playing with the VP reduce their rate of AAE feature production when engaged in the classroom phase as opposed to the bridge building phase. We have also observed that children mimic aspects of the VP's language, posture and bridge building actions. Given these observations, we are motivated to continue to see if these preliminary observations can be substantiated.

5. ACKNOWLEDGEMENTS

The authors thank Justine Cassell, the members of the ArticuLab, the CollaboLab, Darren Gergle, Dan Jurafsky, Doug Downey, and Okechukwu Chika. This research is graciously supported by grant NSF ALT CNV0059012 from the National Science Foundation.

6. REFERENCES

[1] M. Perie, W. Grigg, and P. Donahue, "The Nation's Report Card: Reading 2005," National Center for Education Statistics (NCES), Ed.: US Department of Education, 2005.

[2] E. M. Crowder, "Telling What They Know: The Role of Gesture and Language in Children's Science Explanations," *Pragmatics and Cognition,* vol. 1, pp. 341-376, 1993.

[3] F. Iacobelli and J. Cassell, "Ethnic Identity and Engagement in Embodied Conversational Agents," in *7th International Conference on Intelligent Virtual Agents*, Paris, France, 2007, pp. 57-63.

[4] Q. Li and M. J. Russell, "An Analysis of the Causes of Increased Error Rates in Children's Speech Recognition," in *ICSLP-2002*, 2002, pp. 2337-2340.

[5] A. Potamianos and S. S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing,* vol. 11, pp. 603-616, Nov 2003.

[6] H. Vilhjalmsson, N. Cantelmo, J. Cassell, N. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A. Marshall, C. Pelachaud, Z. Ruttkay, K. Thórisson, H. van Welbergen, and R. van der Werf, "The Behavior Markup Language: Recent Developments and Challenges," in *7th International Conference on Intelligent Virtual Agents*, Paris, France, 2007, pp. 57 -- 63.

[7] Y. I. Nakano, G. Reinstein, T. Stocky, and J. Cassell, "Towards a Model of Face-to-Face Grounding," in *Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, 2003, pp. 553-561.

[8] M. Maatman, J. Gratch, and S. Marsella, "Natural Behavior of a Listening Agent," in *5th International Conference on Interactive Virtual Agents (IVA)* Kos, Greece, 2005.

[9] L.-P. Morency, I. de Kok, and J. Gratch, "Predicting Listener Backchannels: A Probabilistic Multimodal Approach," in *Intelligent Virtual Agents*. vol. 5208/2008: Springer Berlin / Heidelberg, 2008, pp. 176-190.

[10] D. Jurafsky and J. H. Martin, *Speech and Language Processing (2nd Edition)*: Prentice-Hall, Inc., 2006.

[11] A. Stolcke, "SRILM -- an extensible language modeling toolkit," ICSLP-2002, 2002, pp. 901-904.

[12] S. Patel, W. Bosley, D. Culyba, S. A. Haskell, A. Hosmer, T. J. Jackson, S. J. M. Liesegang, P. Stepniewicz, J. Valenti, S.

Zayat, and B. Harger, "A Guided Performance Interface for Augmenting Social Experiences with an Interactive Animatronic Character," in *2006 AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2006, pp. 72-79.

[13] I. Barakonyi, T. Psik, and D. Schmalstieg, "Agents that talk and hit back: animated agents in augmented reality," in *Mixed and Augmented Reality, 3rd IEEE / ACM International Symposium on Mixed and Augmented Reality (ISMAR'04)*, Arlington, VA, 2004, pp. 141-150.

[14] P. Gebhard and M. Klesen, "Using real objects to communicate with virtual characters," in *International Conference on Intelligent Virtual Agents*, 2005, pp. 99-110.

[15] K. R. Thorisson, "Computational Characteristics of Multimodal Dialogue," in *AAAI Fall Symposium on Embodied Language and Action*, Cambridge, MA, 1995, pp. 102-108.

[16] P. Maes, "Artificial life meets entertainment: lifelike autonomous agents," *Communications of the ACM,* vol. 38, pp. 108-114, 1995.

[17] M. Slater, A. Steed, J. McCarthy, and F. Maringelli, "The influence of body movement on subjective presence in virtual environments.," *Human Factors,* vol. 40, p. 469(9), 1998.

[18] M. E. Foster and J. Oberlander, "Corpus-based generation of head and eyebrow motion for an embodied conversational agent," *Language Resources and Evaluation,* vol. 41, pp. 305-323, 2007.

[19] J. Cassell, M. Ananny, A. Basu, T. Bickmore, P. Chong, D. Mellis, K. Ryokai, J. Smith, H. Vilhj, lmsson, and H. Yan, "Shared reality: physical collaboration with a virtual peer," in *CHI '00 extended abstracts on Human factors in computing systems* The Hague, The Netherlands: ACM, 2000.