A Storytelling Robot: Modeling and Evaluation of Human-like Gaze Behavior

Bilge Mutlu
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA, USA
bilge@cs.cmu.edu

Jessica Hodgins
Robotics Institute and
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA, USA
jkh@cs.cmu.edu

Jodi Forlizzi
Human-Computer Interaction Institute and
School of Design
Carnegie Mellon University
Pittsburgh, PA, USA
forlizzi@cs.cmu.edu

***Abstract*— Engaging storytelling is a necessary skill for humanoid robots if they are to be used in education and entertainment applications. Storytelling requires that the humanoid robot be aware of its audience and able to direct its gaze in a natural way. In this paper, we explore how human gaze can be modeled and implemented on a humanoid robot to create a natural, humanlike behavior for storytelling. Our gaze model integrates data collected from a human storyteller and a discourse structure model developed by Cassell and her colleagues for humanlike conversational agents [1]. We used this model to direct the gaze of a humanoid robot, Honda's ASIMO, as he recited a Japanese fairy tale using a pre-recorded human voice. We assessed the efficacy of this gaze algorithm in an experiment that was motivated by results in the literature on human-human communication. Those results suggest that more frequent gaze toward the listener should have a positive affect. We manipulated the frequency of ASIMO's gaze between two participants and used pre and post questionnaires to determine that participants evaluated the robot more positively and performed better in a recall task when ASIMO looked at them more. This experiment provides an evaluation of our gaze algorithm and also adds to the growing evidence that there are many commonalities between human-human communication and human-robot communication.**

I. Introduction

Of the many applications proposed for robots with a human form, education and entertainment are two of the most promising and also likely two with the greatest potential to be successful near term. Entertainment, in particular, can often be scripted, reducing the need for sensing and robustness to changes in the environment. Education, at least in controlled settings such as museums, has similar characteristics. Our research on humanoid robots describes a framework of five design variables - gaze, gesture, proximity to a human partner, small movement, and speech and sound - that feature strongly in the interaction design of a compelling human-robot education and entertainment interactions.

In a storytelling application, all design variables will need to be scripted to act together in a natural manner. In our target application, storytelling, we have first chosen to focus on gaze. This requirement means that the robot should look at the members of the audience, and augment gaze with appropriate gestures while telling the story.

Building on results in the literature for avatar gaze [1] and coding of the actions of a professional storyteller, we implemented a gaze and gesture algorithm for Honda's humanoid robot ASIMO (figure 1) using a combination of hand-coded and automated procedures. The gaze algorithm was automatic, and was based on a hand-coded script of the structure of the written story. A set of generic gestures were added automatically and then supplemented by a few specialized gestures that were scripted by hand to correspond to content in the story.

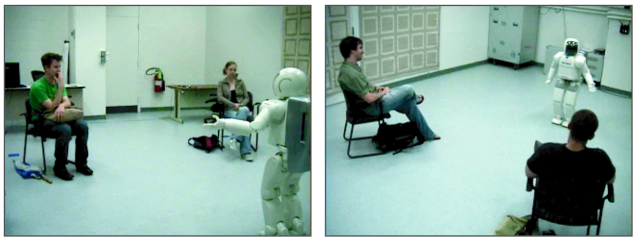

Fig. 1. ASIMO telling a Japanese fairy tale to two listeners.

Gaze is an essential component of human-human communication and is often used for communicating syntactic or semantic signals during speech [2] and storytelling in particular. Studies of gaze suggest that people who look more frequently at others are more likely to be judged more favorably [3]. We used this result to assess our automatic gaze generation algorithm by manipulating the percentage of the time that the robot's gaze was directed at each of two subjects during the telling of a story. Our experimental results matched the predictions in the literature as the subjects who were gazed at less did indeed feel less positively about the robot. They also performed more poorly on a recall test about the story indicating that gaze is also important for comprehension and education. This experiment demonstrated that our model was sufficient to reproduce the effect of at least one aspect of human gaze with a humanoid robot. It also provides further evidence that there are many commonalities between human-robot and human-human communication and that we might do well to explore what is known about human-human communication as we develop humanoid robots that are more interactive and more effective communicators.

through interaction design, maybe this comes at the end.

II. Background

In constructing our experimental hypotheses and design, we build on the existing literature about the social function of gaze in human-human communication. In the next section, we describe results that primarily apply to oratory gaze behavior because our application is storytelling. In the following section, we describe existing models for implementing gaze behaviors in computer agents and avatars.

A. Social Function of Gaze Behavior

Gaze is arguably the most important mechanism humans have for expressing their intentions and understanding the intentions of others. Argyle and Cook argue that gaze serves three main functions in face-to-face interaction: information seeking, controlling the flow of the conversation, and sending syntactic or semantic signals that accompany speech [2]. This last function is the most relevant to our application of storytelling.

When gaze is used for seeking information, attention is directed toward a source in order to read visual signals from our environment. For example, listeners spend most of their time looking at a speaker to supplement auditory information, while speakers spend much less time looking at listeners to reduce the amount of cognitive processing that is done during speech planning [2]. Direction of gaze is affected by the structure and content of the utterance [4]. For example, people look less when they attempt to discuss a cognitively difficult topic [5]. We found that our professional storyteller also spent time gazing away from her listeners and included that in our model of gaze. Gaze also supports speech in communicating syntactic signals such as verbal utterances and emphasis [6] and it is this role that we attempted to emphasize in our story telling application.

In addition to supporting information-seeking and expressing intention, gaze also serves critical social functions. First, gaze patterns communicate interpersonal attitude or affect between speaker and listener. In general, people who look more at others tend to be perceived more favorably, as more competent, friendly, credible, assertive, and socially skilled [3]. Gaze patterns also communicate liking and status among members of a group. For example, in group settings, people tend to look more at group members whom they like [5]. People are observed to look less at others of lower status [7]. Gaze patterns can communicate a speaker's attitude. Speakers tend to gaze at listeners more when they intend to be more persuasive, deceptive, ingratiating, or assertive [3]. Gaze patterns can also communicate the social characteristics of an individual. For example, extroverts are found to look more in general than introverts [8].

Gender can also have a significant effect on gaze behavior. In general, women engage in eye contact more than men do [2], and are shown to look more while listening if they like the speaker. Conversely, men look more while speaking if they like the listener [5], [9].

B. Simulating gaze behavior in agents and robots

Conversational agents have been built that model human-like gaze behavior in order to build simulations of gaze behavior in human-computer conversations [10], [1], [11]. These models include such elements of human-human communication as how speakers look away from listeners at the beginning of an utterance, and toward listeners at the end of an utterance.

Gandalf, an autonomous computer agent, used gaze to display basic attentional cues (e.g. gazing at and turning head to the area of interest) [10]. Other applications have associated a predefined set of gaze behaviors with verbal and thematic markers [12]. For example, when the character said "Let me think..." it also looked up to indicate that cognitive processing was taking place. Peters and O'Sullivan implemented a bottom up gaze behavior that used vision, attention, gaze generation, and memory to allow an agent to pay attention to salient parts of its surroundings [13].

While modeling and simulation of human-like gaze behavior has been explored with conversational agents, less work has examined how gaze might effect social behavior for embodied robots. Sidner and colleagues developed a head turning application that attends to users or objects in an environment and implemented it on Mel, a penguin robot [14]. Nagai implemented a gaze tracking/directing model on the Infanoid robot for experiments in joint attention [15].

III. Hypotheses

Drawing from these results in the social science literature, we formulated two hypotheses about responses to ASIMO's gaze behavior:

Hypothesis 1. Participants who are looked at more will perform better in the recall task than the participants who are looked at less.

Hypothesis 2. Participants who are looked at more will evaluate ASIMO more positively than the participants who are looked at less.

IV. Method

We designed a storytelling experience where ASIMO told a Japanese fairy tale, "The Tongue-Cut Sparrow" [16] to two listeners using a pre-recorded voice. To do so, we developed a human-like gaze model for ASIMO, creating and implementing an algorithm that dynamically directs the robot's gaze based on a coding of the story.

A. Modeling of human-like gaze behavior

Our gaze model is an extension of a model published by Cassell and colleagues [17] with parameters determined by coding the performance of a professional storyteller. Cassell and colleagues developed an empirical model of gaze behavior during turn-taking and within a turn based on the structure of the information conveyed by the speaker [17]. Their model follows the English sentence structure suggested by Halliday [18], who describes the two main structural components of an utterance using the terms "theme" and "rheme." The theme refers to the part of an utterance that sets the tone

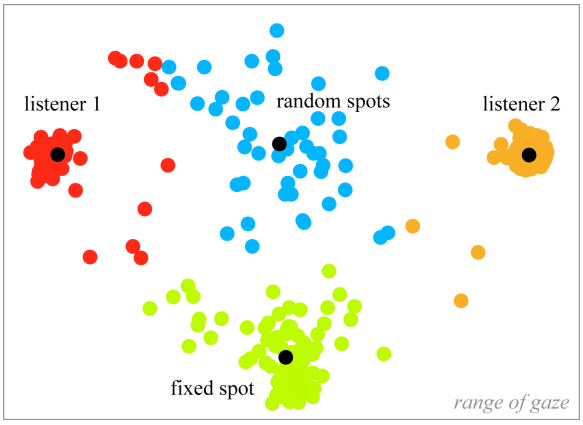

Fig. 2. Clustering of the four gaze locations used by the storyteller.

of the utterance and connects the previous utterance to the next one. The rheme contains the new information that the utterance intends to communicate. For instance, in the sentence "In the evening the old man came home." "In the evening the old man" is the theme while "came home" is the rheme of the utterance. In their model, speakers look away from their listener at the beginning of a theme with 0.70 probability and look at their listeners at the beginning of a rheme with 0.73 probability. They suggested the following algorithm to simulate natural gaze behavior ($distribution(x)$ refers to a randomized function that returns true with probability x). For each proposition:

if proposition is theme **then**
  if beginning of turn or $distribution(0.70)$ **then**
    attach a look-away from the hearer
  end if
else if proposition is rheme **then**
  if end of turn or $distribution(.73)$ **then**
    attach a look-toward the hearer
  end if
end if

We used empirical data collected from a professional storyteller to determine locations and frequencies for the look-away gaze. We videotaped a professional storyteller relating two stories to a two person audience. We used 30 minutes of video data to analyze where and how long the storyteller directed her gaze. The analysis showed that the storyteller gazed at four different kinds of locations: the two members of the audience, a fixed spot on the table in front of her, and a set of random locations in the room. Figure 2 shows a k-means clustering of these four locations with cluster centers in black.

We defined "looking at" as keeping ASIMO's gaze on one listener once it was fixated there. "Looking away" meant looking at the other listener or looking at a random spot or the fixed location. When the gaze was not currently directed at a listener, "looking at" meant looking at one of the listeners, while "looking away" meant looking at any four of the targets with predetermined probabilities. These probabilities were derived from an analysis of the frequencies of our storyteller's

	Listener 1	Listener 2	Fixed spot	Random spot
Frequency (%)	13	11	38	38
Length (%)	38	27	30	5
Min (ms)	477	484	242	360
Max (ms)	15,324	5,914	13,674	4,383
Mean (ms)	2,400	2,262	2,640	1,072

TABLE I
LENGTH AND DISTRIBUTIONS OF GAZE AT EACH LOCATION

gaze at each location. The lengths of gaze at each location followed a normal distribution, which we used to determine the length of the simulated gaze. Table I shows these values for each gaze location.

B. Implementation

This gaze model was used with a human-coded script of the information structure of the fairy tale to simulate human-like gaze behavior. The script marked the start of each theme and rheme and pauses between utterances. Below is the pseudo-code for the algorithm where $probability(x)$ produces a uniform randomized function that returns true with probability derived from the frequencies from our empirical data and $length(x)$ generates a length for the gaze over a normal distribution with mean and standard deviation values generated by our empirical results ($Normal(Mean(x), StDev(x))$).

for each part of the utterance (theme/rheme/pause) **do**
  while the duration of the part **do**
    if current part is pause **then**
      if distribution($probability(random)$) **then**
        gaze at random spot with $length(random)$
      else
        gaze at random spot with $length(fixed)$
      end if
    else if current part is theme **then**
      if distribution($probability(0.70)$) **then**
        if distribution($probability(random)$) **then**
          gaze at random spot with $length(random)$
        else
          gaze at random spot with $length(fixed)$
        end if
      else
        if distribution($probability(listener1)$) **then**
          gaze at random spot with $length(listener1)$
        else
          gaze at random spot with $length(listener2)$
        end if
      end if
    else if current part is rheme **then**
      if distribution($probability(0.73)$) **then**
        if distribution($probability(listener1)$) **then**
          gaze at random spot with $length(listener1)$
        else
          gaze at random spot with $length(listener2)$
        end if
      else
        if distribution($probability(random)$) **then**

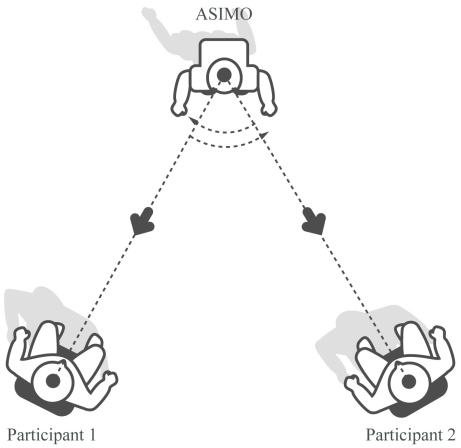

Fig. 3. Experimental setup.

gaze at random spot with $length(random)$
else
gaze at random spot with $length(fixed)$
end if
end if
end if
end while
end for

The gaze algorithm was implemented on ASIMO by following a human-coded script of the story and synchronizing ASIMO's gaze behavior with a pre-recorded voice. Simple arm gestures were automatically added for long utterances. Special gestures such as bowing or acting angry were added by hand when they were semantically appropriate. The location of the participants was not sensed but was determined by placing two chairs at known locations and programming ASIMO to look in that direction. The initiation of the robot's movement was controlled by the experimenter. The robot then introduced himself to the participants, told his story, and then ended the interaction.

C. Evaluation

We conducted a between-subjects experiment where participants listened to ASIMO while he told a Japanese fairy tale. We manipulated ASIMO's gaze behavior to gaze at one of the participants with 20% frequency and the other participant with 80% frequency. Participants were placed at the same distance from ASIMO and space was left between them so that they would not interact with each other and the robot's gaze at each participant would be easily distinguishable (Figure 3).

a) Experiment procedure.: Participants were first given a brief description of the experiment procedure. After the introduction, participants were asked to answer a pre-experiment questionnaire. This was followed by providing the participants with more detail on their task. ASIMO then introduced himself and performed the storytelling task. After listening to ASIMO's story, participants performed a distractor task, where they listened to another story on tape ("The Flying Trunk" by Hans Christian Andersen [19]. Before listening to either of the stories, they were told that they would be asked questions regarding one of these stories. All participants were asked questions regarding ASIMO's story. After completing the task, participants answered a post-experiment questionnaire regarding their affective state, their perceptions of the robot, and their demographic information. ASIMO's story, the story on tape, and the whole experiment took an average of 17.5 minutes, 7.5 minutes, and 35 minutes respectively. The experiment was run in a dedicated space with no outside distraction. A male and a female experimenter were present in the room during the experiment. All participants were paid a $10 in compensation for their participation.

b) Measures and sample.: All factors in the experiment were identical for each participant except for the two controlled factors: the frequency of the robot's gaze at each participant (a manipulated independent variable) and the participant's gender (a measured independent variable). The dependent variables measured were task performance, the participant's own affective state, their positive evaluation of the robot, their perceptions of the robot's physical, social, and intellectual characteristics, their involvement in and enjoyment of the task, and participant demographics. The post-experiment questionnaire included a question as a manipulation check, "How much did the robot look at you?". Seven-point rating scales were used for all scales.

Twenty (12 males, 8 females) undergraduate and graduate students from Carnegie Mellon University participated in the experiment. Ten participants were assigned to the "looked at 80% of the time" condition. The other ten participants were assigned to the "looked at 20% of the time" condition. Ages of the participants ranged from 19 to 33. Participants were chosen to have a variety of majors including management sciences, social sciences, art, and engineering.

V. Results

Our data analysis used three methods; repeated measures analysis of variance (MANOVA), regression (Least Squares Estimation), and multivariate correlations. The first method applied an Omnibus F-Test to see if the difference between pre-experiment and post-experiment measurements was significant across the two experiments, task structures, and/or genders. The second technique used a linear regression on the variables that were significant across conditions to identify the direction of main effects and interactions. The last method looked at how these variables correlated with each other. We also ran reliability tests and factor analysis on the scales we used for measurement.

Item reliabilities for all partner (robot), task, and self evaluation scales except the mutual liking scale ($\alpha = 0.54$) were high. We also ran a factor analysis of all the items that we used for partner evaluation and created a highly reliable ($\alpha = 0.91$), 8-item scale for partner positive evaluation. An analysis of the manipulation check showed that the participants were aware that they were looked at more or less by the robot (F[1:16]=3.48, p¡0.01).

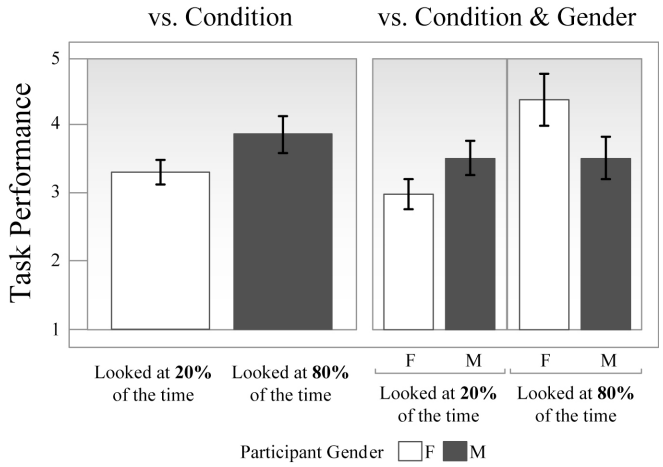

Fig. 4. Main effect of condition and interaction between condition and participant gender on task performance.

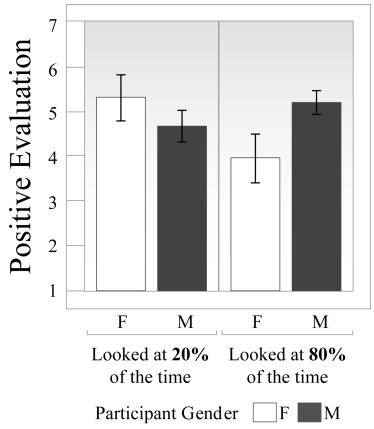

Fig. 5. Interaction between condition and participant gender on positive evaluation of the robot.

Consistent with our first hypothesis, a regression on the performance measure showed that participants who were looked at more performed significantly better in the recall task (answering questions regarding ASIMO's story) than those who were looked at less (F[1:16]=5.15, p=0.03). When participant's gender was included in the statistical model, the effect was significant only in females (F[1:16]=8.58, p¡0.01) while men did not show any significant difference across conditions (Figure 4).

Our analysis of the ratings of the positive evaluation scale showed no significant main effect but a significant interaction of experimental condition and participant gender (Figure 5). Men rated ASIMO more positively when they were looked at more while women's evaluations were higher when they were looked at less (F[1:16]=5.62, p=0.03). Although this result reveals significant interactions with participant's gender, it is not consistent with the prediction in our second hypothesis. Analysis of scales of participant's affect, task enjoyment, and task involvement did not show any significant effects or interactions.

We also looked at how our scales correlated with participant's computer use, their familiarity with robots, and video gaming experience. A multivariate analysis using Pearson's correlation coefficient showed that ratings of the positive evaluation scale was highly correlated with video gaming experience (r=0.65, p¡0.01), while not correlated with computer use or familiarity with robots. This correlation held for both genders although it was stronger in men. Video gaming experience was also high correlated with task enjoyment (r=0.53, p=0.02).

VI. Discussion

Our results supported the first hypothesis: the frequency of the robot's gaze affected performance on the recall task. These results have design implications for human-robot communication, particularly in education or other applications where important material is being conveyed. For example, a humanoid might try to engage a particular listener by looking at that listener more when he/she does not appear to be attending. In the future, human-robot interactions might be designed so as to improve the recall of the material being presented.

The second hypothesis, that participants who are looked at more will evaluate the robot more positively, was also supported. However, when we included gender as a variable in that analysis, we found that women liked the robot more when they were looked at less. This result was surprising as it was not predicted by the existing literature in human-human communication. One potential explanation is that although ASIMO's gaze behavior was quite human-like, it was not natural or rich enough. Further experiments, on robots with both richer and poorer expressive capabilities and perhaps with a performance that is completely choreographed for a particular story might help to explain this result.

We also found that positive evaluations of ASIMO were highly correlated with participant's video gaming experience and not with their computer use, which suggests that people might perceive ASIMO as more like a video-game character or avatar than like a computer. This result suggests that we should rely most heavily on results in the interaction literature for computer agents rather than computers themselves when we design interactive experiences with humanoid robots.

Some elements of the professional storyteller's gaze were not accounted for by our model. For example, she occasionally switched from looking at one listener to looking at the other listener during a theme or rheme, but we could not find a pattern on which to model this behavior. Although we believe that our gaze model was sophisticated enough not to be perceived as algorithmic by the participants, it is possible that the introduction of more complexity based on more detailed coding of human performances would improve its naturalness. We plan to gather more data and use it for the next iteration of our gaze model.

Although we were careful to make our gaze model as human-like as possible, there were still some unnatural elements to ASIMO's story telling performance. For example,

ASIMO's arm gestures were found distracting by some participants, perhaps because of the motor noise that occurs during the robot's arm movements. Another possible explanation is that our library of gestures is too limited forcing most of the gestures to be "generic" motions of the arms to the side or front of the robot. A human storyteller would likely use gestures that were more closely matched to the content of the particular story. Some subjects reported that ASIMO's story was too long (17.5 minutes) and it might be easier to create a compelling performance for a shorter story.

Our experiment was aided by some wizard-of-oz steps in that ASIMO did not sense where his audience was seated or allow responses from them during the telling of the story. Robust vision and natural language techniques would be required to address these issues and allow the construction of a truly interactive experience for the participants.

ACKNOWLEDGMENTS

This work was made possible by the financial support from the NSF under IIS-0121426 and ECS-0325383 and an equipment loan from Honda R&D Co., Ltd. to Takeo Kanade and the second author. We would like to thank our storyteller Gunnhildur Jonsdottir for allowing us to model her gaze.

REFERENCES

[1] J. Cassell, C. Pelachaud, N. I. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone, "Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents," *Computer Graphics*, 1994.

[2] M. Argyle and M. Cook, *Gaze and Mutual Gaze*. Cambridge University Press, 1976.

[3] C. L. Kleinke, "Gaze and eye contact: A research review," *Psychological Bulletin*, pp. 78–100, 1986.

[4] A. Kendon, "Some functions of gaze direction in social interaction," *Acta Psychologica*, vol. 32, pp. 1–25, 1967.

[5] R. V. Exline and L. C. Winters, *Affect Relations and Mutual Gaze in Dyads*, S. Tomkins and C. Izzard, Eds. Springer, 1965.

[6] N. Chovil, "Discourse-oriented facial displays in conversation," *Research on Language and Social Interaction*, vol. 25, pp. 163–194, 1992.

[7] A. Mehrabian, "Inference of attitudes from the posture, orientation, and distance of a communicator," *Journal of Consulting and Clinical Psychology*, vol. 32, pp. 296–308, 1968.

[8] N. A. Mobbs, "Eye-contact in relation to social introversion-extraversion," *British Journal of Social Clinical Psychology*, vol. 7, pp. 305–306, 1968.

[9] A. Kendon and M. Cook, "Consistency of gaze patterns in social interaction," *British Journal of Psychology*, vol. 60, pp. 481–494, 1969.

[10] K. R. Thorisson, "Communicative humanoids: A computational model of psychosocial dialogue skills," Ph.D. dissertation, Massachusetts Institute of Technology, 1996.

[11] C. Pelachaud, N. I. Badler, and M. Steedman, "Generating facial expressions for speech," *Cognitive Science*, vol. 20, pp. 1–46, 1996.

[12] C. Pelachaud and M. Bilvi, "Modeling gaze behavior for conversational agents," in *Proceedings of the International Working Conference on Intelligent Virtual Agents*, 2003.

[13] C. Peters and C. OSullivan, "Bottom-up visual attention for virtual human animation," *Computer Animation and Social Agents*, vol. 2003, pp. 111–117, 2003.

[14] C. Sidner, C. Lee, C. Kidd, and N. Lesh, "Explorations in engagement for humans and robots," in *Proceedings of the International Conference on Humanoid Robots*, 2004.

[15] Y. Nagai, "Joint attention in infant-like robot based on head movement imitation," in *Proceedings of the International Symposium on Imitation in Animals and Artifacts*, 2005.

[16] Y. T. Ozaki, *The Japanese Fairy Book*. Tokyo: Tuttle Publishing, 1970.

[17] J. Cassell, O. E. Torres, and S. Prevost, *Turn Taking vs. Discourse Structure: How Best to Model Multimodal Conversation*. Kluwer, 1998.

[18] M. Halliday, *Intonation and Grammar in British English*. Mouton, 1967.

[19] H. C. Andersen, *Tales. Vol. XVII, Part 3, The Harvard Classics*. P.F. Collier & Son, 190914; Bartleby.com, 2001.