Evaluating Models of Speaker Head Nods for Virtual Agents

Jina Lee, Zhiyang Wang and Stacy Marsella
Institute for Creative Technologies
University of Southern California, USA
{jlee, zwang, marsella}@ict.usc.edu

ABSTRACT

Virtual human research has often modeled nonverbal behaviors based on the findings of psychological research. In recent years, however, there have been growing efforts to use automated, data-driven approaches to find patterns of nonverbal behaviors in video corpora and even thereby discover new factors that have not been previously documented. However, there have been few studies that compare how the behaviors generated by different approaches are interpreted by people. In this paper, we present an evaluation study to compare the perception of nonverbal behaviors, more specifically head nods, generated by different approaches. Studies have shown that head nods serve a variety of communicative functions and that the head is in constant motion during speaking turns. To evaluate the different approaches of head nod generation, we asked human subjects to evaluate videos of a virtual agent displaying nods generated by a human, by a machine learning data-driven approach, or by a hand-crafted rule-based approach. Results show that there is a significant effect on the perception of head nods in terms of appropriate nod occurrence, especially between the data-driven approach and the rule-based approach.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Intelligent agents

General Terms

Experimentation, Human Factors

Keywords

Virtual Agents, Embodied Conversational Agents, Nonverbal Behaviors, Head Nods, Evaluation

1. INTRODUCTION

During face-to-face interaction, we use not only verbal behaviors but also nonverbal behaviors to deliver our communicative intents. These nonverbal behaviors serve to repeat, contradict, substitute, complement, accent, or regulate spoken communication [11]. During speaking turns, we make various arm gestures, facial expressions, or posture shifts, but our head is also in constant motion. Research on nonverbal behaviors has identified a number of important communicative functions served by these head movements [6] [7] [9] [17]. Nods may be used to show our agreement with what the other is saying, to emphasize a certain point, or to request backchannels from the listener. We may also shake our heads to express disapproval and negation, or tilt our heads upward along with gaze aversion when pondering something. Head movements are also influenced by our emotions. For example, Mignault and Chaudhuri [18] found that a bowed head connotes submission, inferior emotions (i.e., shame, embarrassment, etc.), and sadness, whereas a raised head connotes dominance, superiority emotions (i.e., contempt and pride), and happiness. Tom et al. [23] also found that overt head movements can be instrumental in the formation of an observer's affective response.

Following the psychological research identifying the important functions served by head movements, many virtual agent systems have modeled and incorporated head movements. The incorporation of appropriate head movements in virtual agents is shown to have positive effects during human-agent interaction. For example, in [20] Munhall et al. found that human subjects classified more syllables correctly when a virtual agent displayed appropriate head movements while talking, compared to when the agent displayed no head movements or distorted movements.

Previously many of the virtual humans used hand-crafted models to generate nonverbal behaviors including head movements. To specify which behaviors should be generated at each given context, the knowledge from the psychological literature is used to construct a set of rules that associate salient factors to certain nonverbal behaviors [4] [5] [12] [14] [2]. However, there are limitations with this approach. One major drawback is that the rules have to be hand-crafted. This means that the author of the rules is required to have a broad knowledge of the phenomena he/she wishes to model. However, as more and more factors are added that may influence the myriad of behaviors generated, it becomes harder to specify how all those factors contribute to the overall outcome. Unless the rule-author has a complete knowledge on the correlations of the various factors, manual rule construction may suffer from sparse coverage of the rich phenomena.

Recently, there have been growing efforts to use corpora of nonverbal behaviors more extensively [19] [3] [10] [1] [15] [16]. In this data-driven approach, often machine learning techniques are used to learn the patterns of behaviors. One

Cite as: Evaluating Models of Speaker Head Nods for Virtual Agents, Jina Lee, Zhiyang Wang and Stacy Marsella, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. XXX-XXX.

of the advantages of this approach is that the process is automated. Although a good understanding of the phenomena is still important, with machine learning approach, it is no longer necessary for the author of the model to have a complete knowledge of the complex mapping between the various factors and behaviors. Another advantage is that this method is flexible and can be customized to learn for a specific context. For example, to learn the head nod patterns of different cultures, we may train separate models with each culture's data. Similarly, if we wish to learn gesture patterns with individualized styles, we can train each model with data from specific individuals. The major disadvantage of this approach is the limited availability of gesture corpora. It takes a long time to gather data and annotate the behaviors suitable to the researcher's purpose. However, there are growing efforts to standardize the annotation scheme and automate the annotation process.

Although many works that model nonverbal behaviors have been evaluated individually, there are considerably fewer studies that compare the different approaches to behavior modeling. Further, we are especially interested in the perception of naturalness of the behaviors realized by different approaches. Therefore, we stress the importance of evaluation studies with human subjects. A model may produce behaviors that are in accordance with the psychological studies or closely match the original human data (and thus producing high F-score values), however, this does not directly guarantee that they will also look natural to human eyes.

The goal of this paper is to compare and evaluate nonverbal behaviors generated by different approaches. To do this, we focus on comparing the speaker head nods generated by two of our previous works [14] and [16]. The two works differ in the approaches for generating nonverbal behaviors; the former is a rule-based system containing a set of nonverbal behavior rules that identify specific factors known to be associated with certain nonverbal behaviors from the psychological literature, whereas the latter is a data-driven approach that learns the patterns of speaker head nods from the gesture corpora. In this paper we present the results of our evaluation study that compares the perception of head nods driven by the two approaches as well as nods made by humans.

The following sections describe the research on identifying the patterns and functions of head movements and modeling them for virtual agents. We then describe our different approaches for modeling head nods and the evaluation study with human subjects. Results show that there is a significant effect on the judgement of appropriate head nod occurrence, especially between the data-driven approach and the rule-based approach. Finally, we discuss the results and propose future directions.

2. RELATED WORK

The functions and patterns of head movements during face-to-face communication have been studied in various disciplines [6] [7] [9] [17]. Heylen [7] summarizes the functions of head movements during conversations. Some included are: to signal yes or no, enhance communicative attention, anticipate an attempt to capture the floor, signal the intention to continue, mark the contrast with the immediately preceding utterances, and mark uncertain statements and lexical repairs. Kendon [9] describes the different contexts in which the head shake may be used. Head shake is used with or without verbal utterances as a component of negative expression, when a speaker makes a superlative or intensified expression as in 'very very old,' when a speaker self-corrects himself, or to express doubt about what he is saying. In [17], McClave describes the linguistic functions of head movements observed from the analysis of videotaped conversations; lateral sweep or head shakes co-occurs with concepts of inclusivity such as 'everyone' and 'everything' and intensification with lexical choices such as 'very,' 'a lot,' 'great,' 'really.' Side-to-side shakes also correlate with expressions of uncertainty and lexical repairs. During narration, head nods function as signs of affirmation and backchannel requests to the listeners. Speakers also predictably change the head position when discussing alternatives or items in a list.

In accordance with the studies on nonverbal behaviors, many virtual agents model these behaviors to realize their communicative functions. Some generate the behaviors according to the discourse structure or 'conversation phenomena.' For example, REA's [4] verbal/nonverbal behaviors are designed in terms of conversational functions, where it employs head nods to send feedbacks and head toss to signal openness to engage in conversations. BEAT [5] generates eyebrow flashes and beat gestures when the agent describes a new object that is part of the rheme in the discourse structure of the utterance. [2] is a system for automatic non-verbal generation in which head nod is used as a basic gesture type for listener or is used when no other specific gesture can be suggested. In our previous work, we [14] developed a rule-based system that analyzes the syntactic and semantic structure of the surface text to extract the salient factors and associate them with various nonverbal behaviors.

Other virtual agents focus on generating expressive behaviors according to the agent's emotional state. Deira [12] is a reporter agent that generates basic head movements (including facial expressions) at fixed intervals but also produces more pronounced movements as the agent's excitement rises during the report. Similarly, ERIC [22] is a commentary agent that shows 'idle' gestures when no other gestures are requested, but generates various nonverbal behaviors according to its emotional state. Busso et al. [3] use audiovisual signals to synthesize emotional head motion patterns. They use prosodic features and facial expressions recorded from human speakers to build models for each emotional category and use them to synthesize head motions illustrated through an animated face. Their evaluation shows that head motion modifies emotional perception of facial animation especially in valence and activation domain.

Increasingly, researchers are using various gesture corpora and applying automated methods to find regularities in nonverbal behaviors. Morency et al. [19] use a corpus of human-human interaction and create a model that predicts listener's backchannel head nods using the speaker's multimodal features (e.g. prosody, spoken words, eye gaze). Kipp et al. [10] perform a data-driven approach to generate hand and arm gestures with individualized styles and introduce the concept of 'gesture units' that produce more continuous flow of movement. Similarly, Bergmann and Kopp [1] use a corpus of speech and direction-giving and landmark description gestures to learn Bayesian Decision Networks to model the generation of iconic gestures. We [15] also used machine learning techniques on gesture corpora to predict when speaker head nods should occur. We focused on using linguistic features to learn hidden Markov models and the re-

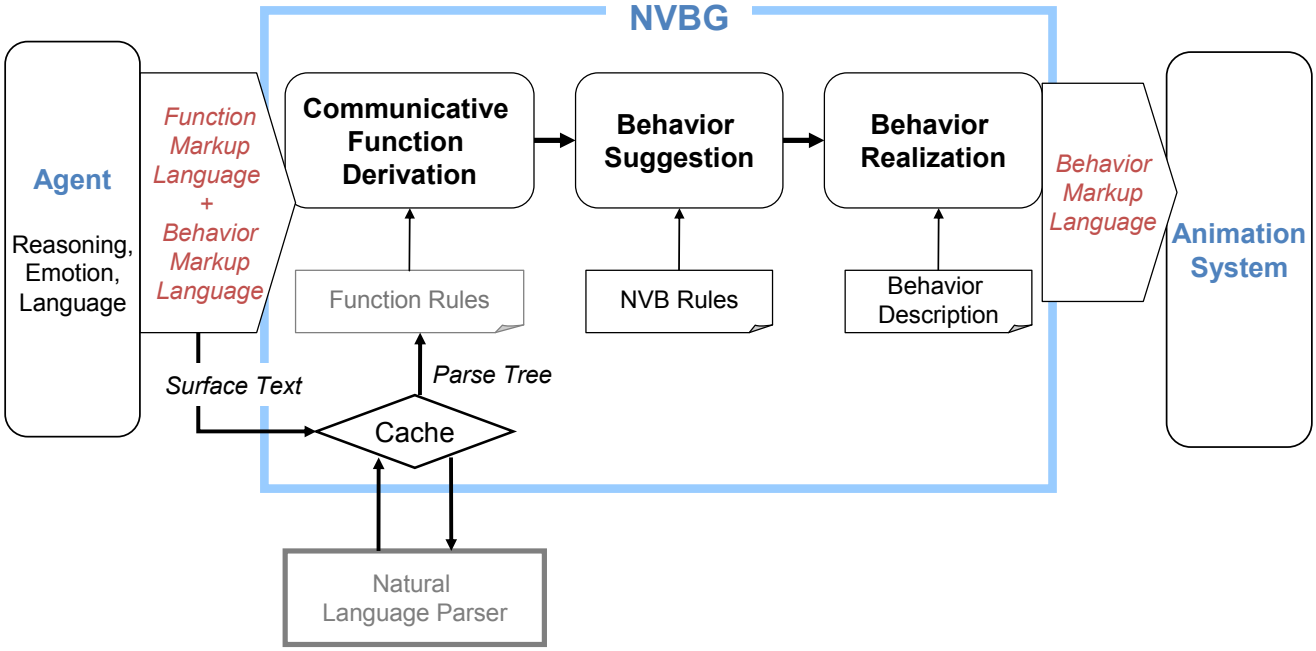

Figure 1: Architecture of the Nonverbal Behavior Generator (Rule-based Approach)

sults show that they were able to predict speaker head nods with high precision, recall and F-score rates even without a rich markup of the surface text. We also showed that using affective sense of each utterance can improve the learning [16].

3. MODELING SPEAKER HEAD NODS

In this section, we describe two different approaches for generating speaker head nods that we later evaluate. First we describe the rule-based approach used in the Nonverbal Behavior Generator [14] that generates head nods along with other behaviors, then the data-driven approach to learn hidden Markov models that predict speaker head nods [16].

3.1 Rule-based Approach

The Nonverbal Behavior Generator (NVBG) [14] is a tool that automates the selection and timing of nonverbal behaviors for virtual agents. It uses a rule-based approach that generates behaviors given the information about the agent's cognitive processes but also by inferring communicative functions from a surface text analysis. The rules within NVBG were crafted using psychological research on nonverbal behaviors as well as the our own study of corpora of human nonverbal behaviors to specify which nonverbal behaviors should be generated at each given context. In general, it realizes a robust process that does not make any strong assumptions about markup of communicative intent in the surface text. In the absence of such markup, NVBG can extract information from the lexical, syntactic, and semantic structure of the surface text that can support the generation of believable nonverbal behaviors.

Fig. 1 shows the architecture of NVBG. The information on the agent's communicative intents, emotional state and surface text is passed from the agent's cognitive module to NVBG in the form of Function Markup Language (FML) and Behavior Markup Language (BML) [13]. NVBG then sends the surface text to the natural language parser to obtain the parse tree that tells NVBG the syntactic structure of the utterance. In the *Communicative Function Derivation* stage, NVBG derives the communicative function from the information both given and derived through its analysis on the surface text and additional information given by analyzing the syntactic and semantic structure of the surface text. Some examples of communicative functions include affirmation, inclusivity, intensification, etc. (see [14] for details). In the *Behavior Suggestion* stage, a set of nonverbal behavior rules maps the derived communicative functions to various behaviors, which then gets specified through Behavior Markup Language in the *Behavior Realization* stage. NVBG can be incorporated with any animation system that can process BML.

As mentioned above, the nonverbal behavior rules were defined from the psychological research on nonverbal behaviors and additional video analysis. We defined the mappings of the rules between the communicative functions and specific behaviors which were found in the research studies. The video data analysis was conducted to validate the findings in the literature as well as to define the dynamic properties of the behaviors including speed, repetition and span of behaviors (e.g. word-level, phrase-level, or cross-syntactical boundaries). Additional mappings between communicative functions and behaviors were also observed and constructed as nonverbal behavior rules. For example, interjections were observed to occur usually with a fast, large magnitude of nod. In addition, the priority values of the nonverbal behavior rules were also defined from the video analysis for cases where one or more rules could overlap at the same utterance segment.

3.2 Data-driven Approach

We also used a data-driven approach using machine learning techniques to build a domain-independent model of speaker head nods [15] [16]. To promote the reusability of the model, we focused on using linguistic features that are easily obtainable across different virtual human systems such as part of speech tags, phrase boundaries, and dialog acts.

The overview of the head nod prediction framework is shown in Fig. 2. To construct the data set, from the gesture corpus we obtained the head movement labels, which serve as the ground truth of when speakers displayed different head movements, as well as the transcripts of each speaker and dialogue acts of each utterance in the corpus. Similar to the NVBG work, we processed each utterance through a natural language parser to obtain information on the syntactic structure of the text. Additionally, we detected affective sense of each utterance by processing it through Affect Analysis Model [21]. Finally, we looked for key lexical entities shown to be associated with head nods from psychological research, such as *yes, very*, and *quite*.

Once we encoded and aligned the features, we selected a subset of those features most correlated with head nods. Feature selection process was performed to reduce the number of parameters to learn for the particular type of model they trained (hidden Markov Model); adding another feature means one needs more data samples to learn how the combinations of the features affect the outcome. Therefore, with a limited amount of data, it is necessary to keep the number of features low.

To learn the model, we trained a hidden Markov model with data samples that accompany head nods. Once the models are learned, a new sample data can be passed through the model to predict whether or not a head nod should occur. The evaluation of the learned model was measured through accuracy, precision, recall, and F-score rates and the results show that the model achieved high measurement values. This shows us that we were able to predict speaker head nods even when using only those features obtained through shallow parsing of the surface text. Additionally, we showed that using information on the affective sense of the utterance improves the learning compared to when no affective information was used.

Although the results show that we were able to train good head nod models without a rich markup of the text, the evaluation was done mathematically. Instead, what should be emphasized is whether the generated head nods look natural or appropriate to *humans*. Despite the fact that the data-driven model can generate head nods with high precision, recall, or F-score rates, it could generate one nod at a wrong timing, which could look absurd to the human user interacting with the virtual agent. The next section describes our evaluation study with human subjects comparing the head nods generated by rule-based approach and data-driven approach.

4. EVALUATION STUDY

An online evaluation study was conducted to compare the perception of speaker head nods generated by the models described in the previous section. Our main interest is to investigate how natural the different head nods are perceived by humans and compare them to head nods made by real humans but displayed through a virtual agent. To answer this, we compare the following schemes for generating head nods:

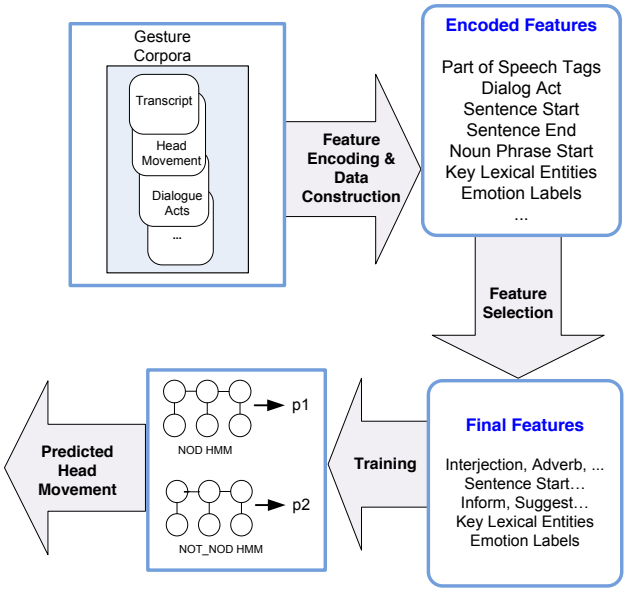

Figure 2: Overview of Speaker Head Nod Model (Data-driven Approach)

- Head nods made by humans but displayed through a virtual agent
- Head nods generated by a data-driven approach and displayed through a virtual agent
- Head nods generated by a rule-based approach and displayed through a virtual agent

In this study, we hypothesize that

> *Nods generated by a data-driven approach will be perceived to be more natural than nods generated by rule-based approach.*

We base the hypothesis from the fact that because data-driven approach uses corpus on real human data to model nods, this approach may capture the 'naturalness' better than rule-based approach. On the other hand, while it may be easier to explain the communicative functions of each nod generated by the rule-based approach, at times the nods may look unnatural since one or two nods could be used to realize a communicative function that spans over several phrases or utterances. In addition to the main hypothesis, we also expect that the human-made nods will look more natural than nods from the data-driven approach or rule-based approach.

4.1 Methods

Participants

37 participants were recruited via email and web postings. There were 19 males and 18 females with ages ranging from 19 to 41 (M=28.4 years, SD=6.47 years).

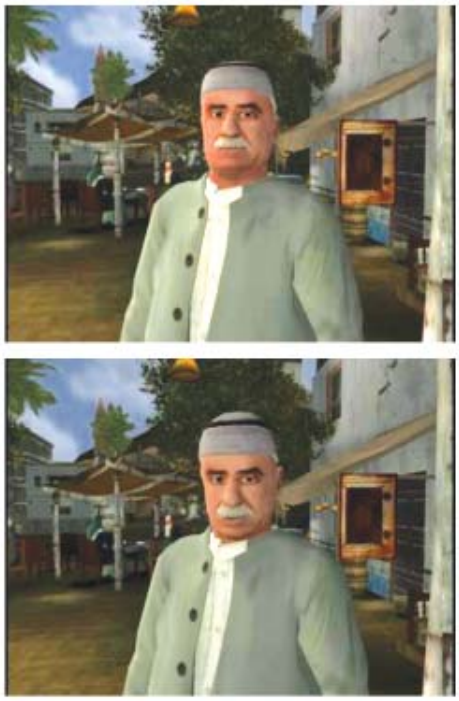

Figure 3: Snapshot of the video clip shown in the evaluation study

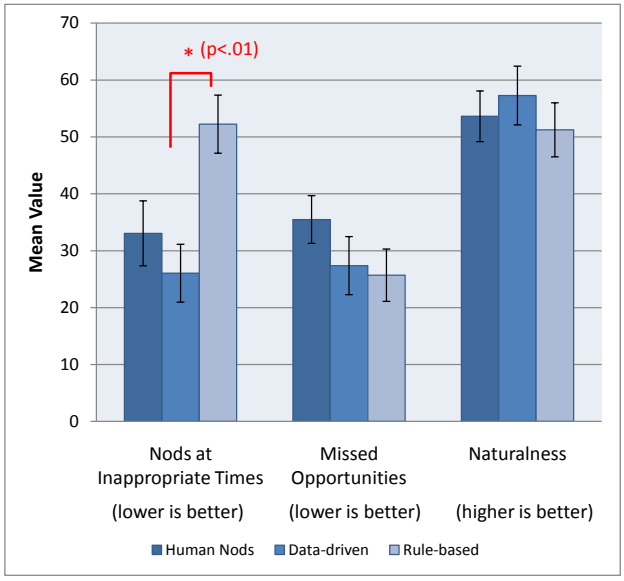

Figure 4: Mean values for the evaluation study. Vertical bars denote the 95% confidence intervals.

Stimuli

We created video clips of a virtual agent displaying head nods while speaking an utterance. Fig. 3 shows a snapshot of the video clip. We randomly selected 7 utterances from the gesture corpus used in the data-driven approach. None of the utterances were used during the training process for learning the speaker head nod models. We then passed these utterances through both the rule-based model (NVBG) and data-driven model to obtain head nod predictions. With the nod predictions, we created three versions of video clips for each utterance: head nods displayed by human in the gesture corpus, head nods generated by data-driven approach, and head nods generated by rule-based approach. Therefore, there were a total of 21 video clips (7 utterances x 3 conditions). In all three conditions, the magnitude, velocity, and length of the nods were unified; the models only predicted the timing of the nods, not the dynamics. Therefore, the only differences among the different conditions were the frequency of the nods. No other nonverbal behaviors were generated except for the lip syncing motion and eye blinking. The average numbers of nods in an utterance for human nods, data-driven approach, rule-based approach were 3.14, 5.29, and 4.5.

Design and Procedure

All evaluation studies were completed online. Participants first filled out a demographic questionnaire asking their age, gender, education level, ethnicity, and occupation. In order to compare the naturalness of the different head nods, one set of video clips (i.e. one utterance) was randomly selected which consisted of three clips, representing each condition (human nods, nods from data-driven approach, nods from rule-based approach). The order of these clips was randomized as well. Each video clip lasted about 10 seconds. After watching each video, users were asked to answer questions on the naturalness of the head nod timings and the overall naturalness of the nodding behavior. The specific questions asked were,

1. How often do you think the agent head nods at inappropriate time?
2. How often do you think the agent missed head nod opportunities?
3. How natural is the nodding behavior overall?

Participants answered the questions using a scale from 0 to 100 (0 meaning 'Never' or 'Not natural at all' and 100 meaning 'Always' or 'Very Natural'). The whole survey lasted about 5-10 minutes.

4.2 Results

The analyses of the answers are based on repeated measures ANOVA with modeling approach as within-subject variable. Bonferroni adjustments were used for post-hoc pairwise comparisons. All the analyses were performed with SPSS. Fig. 4 shows the mean values for the three questions.

For the perception of nods at inappropriate times (Question 1), there was a significant effect of the generation approach ($F(2, 72)=6.291$, $p=.003$). In general, participants perceived that nods generated by data-driven approach ($M = 26.054$, $SE = 5.083$) had fewer cases of nods at inappropriate timing, followed by human nods ($M = 33.054$, $SE = 5.705$) and nods generated by rule-based approach ($M = 52.243$, $SE = 5.107$). The pairwise comparisons (see Table 1) shows that there is a significant difference between data-driven approach and rule-based approach.

For the perception of missed head nod opportunities (Question 2), participants generally perceived the rule-based approach missed less nod opportunities followed by data-driven

Comparison		Mean Diff.	Std. Error	Sig.	95% CI	
					Lower Bound	Upper Bound
Human	vs. Data-driven	7.000	7.478	1.000	-11.779	25.779
	vs. Rule-based	-19.19	7.970	.064	-39.203	.824
Data-driven	vs. Human	-7.000	7.478	1.000	-25.779	11.779
	vs. Rule-based	-26.189*	7.476	.004	-44.961	-7.418
Rule-based	vs. Human	19.189	7.970	.064	-.824	39.203
	vs. Data-driven	26.189*	7.476	.004	7.418	44.961

Table 1: Pairwise comparison of mean values of Question 1 (nods at inappropriate times). Significant differences are denoted by * (p<.01)

approach and human nods, but there was no significant effect. For the overall naturalness of the nodding behaviors (Question 3), participants rated the nods generated by data-driven approach the highest followed by the human-made nods and nods generated by rule-based approach. However, similar to the previous question, there was no significant effect.

4.3 Discussion

The results of the evaluation study show that there was a significant effect of the generation approach on the ratings for nods at inappropriate times and furthermore, that there was a significant difference between the data-driven approach and the rule-based approach. Therefore, we can conclude that our hypothesis was partly validated. Additionally, there is a general trend that the nods from data-driven approach does better than the human nods; the data-driven approach produced better results than human nods in all three questions.

There are several possible explanations for the results. First of all, the data-driven model is a general model which is trained on the data from many different people. Therefore, the model predicts where people will most likely nod by analyzing various features, especially in this model, the linguistic features. On the other hand, the videos showing the human nods are based on nodding behaviors of an individual (the individual who originally spoke the utterance), and thus the participants are comparing a particular person's nods to the those from a general model. Since each individual styles of nodding behaviors (and other nonverbal behaviors) differ, the results may be indicating that an average behavior is perceived as more appropriate than nods with an individual style, even though they are a reflection of a real human's nods. The study of Huang et al. [8] also showed a similar phenomenon but in the case of listener backchannels, whereby a model learned from consensus sampling was rated more natural than the backchannel nods made by a human but displayed through a virtual agent.

Secondly, with regards to the third question asking about the overall naturalness of the nodding behavior, instead of only focusing on when nods occurred, participants may also have included other dynamics of the nods (e.g. speed or number of repetitions) or even the display of other behaviors to the evaluation criteria. Since the data-driven model used in this study currently only predicts when the nods should occur and not other dynamics, when we created the video clips we unified the speed, number of repetitions, starting direction (down/up), and the magnitude of nods across different conditions. Therefore, the only difference among the videos shown to the participants was when the nods occurred. However, some of the participants' post-evaluation comments show that they felt the behavior was unnatural because it was missing body movements or hand gestures, or that they felt like they saw some side to side motion. This shows that they might have included other aspects when evaluating the naturalness of nodding behaviors.

5. CONCLUSION

In this paper we presented our evaluation study for comparing speaker head nods generated by different approaches. We focused on comparing the nods generated by humans, data-driven approach, and rule-based approach. We conducted an evaluation study in which we showed three different videos with head nods generated by different approaches and asked participants to evaluate on the appropriateness of the nods. Results show that there was a significant effect of the generation approach on the ratings for nods at inappropriate times and that the data-driven approach had significantly less nods displayed at inappropriate times than the rule-based approach. However, there were no significant differences across the different approaches on the perception of missed nod opportunities or the overall naturalness of the nodding behavior.

This work could be extended in several ways. First of all, in addition to the appropriateness of the nods, we are interested in the implications the subjects get from observing the behaviors. A challenge here would be in figuring out what the right questions are and how to carefully construct them. Secondly, we plan to repeat the study including more sets of videos with longer utterances. This will allow the participants to see the difference between the data-driven approach and the rule-based approach more clearly and remove biases that may come from the individual nodding styles when watching the nods made by humans. Thirdly, we need to find more effective ways to measure naturalness. We need to investigate what factors truly contribute to the naturalness of behaviors and carefully phrase the questions and measuring scales to reduce misinterpretations. Finally, we want to evaluate nonverbal behaviors generated by other models as well. In this study, we only compared one case of data-driven model and rule-based model to human nods. However, there are many different works modeling nonverbal behaviors for virtual agents as listed in Section 2. We want to include these different works in the evaluation study to get a more generalized conclusion of which approach produces more natural-looking behaviors. Furthermore, we may even combine the data-driven and rule-based approaches to investigate if it improves the quality of the behaviors than either of the approaches.

6. ACKNOWLEDGMENTS

This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM), and the content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

7. REFERENCES

[1] K. Bergmann and S. Kopp. GNetic-using bayesian decision networks for iconic gesture generation. In *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, pages 76–89, 2009.

[2] W. Breitfuss, H. Prendinger, and M. Ishizuka. Automated generation of non-verbal behavior for virtual embodied characters. In *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*, pages 319–322, New York, NY, USA, 2007. ACM.

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

[4] J. Cassell. More than just another pretty face: Embodied conversational interface agents. *Communications of the ACM*, 43:70–78, 2000.

[5] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore. BEAT: the behavior expression animation toolkit. In *SIGGRAPH '01: Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 477–486, New York, NY, USA, 2001. ACM.

[6] U. Hadar, T. J. Steiner, and F. C. Rose. Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9(4):214–228, 1985.

[7] D. Heylen. Challenges ahead: Head movements and other social acts in conversations. In *AISB 2005, Social Presence Cues Symposium*, 2005.

[8] L. Huang, L.-P. Morency, and J. Gratch. Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior. In *AAMAS '10: Proceedings of the 9th international joint conference on Autonomous agents and multiagent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2010.

[9] A. Kendon. Some uses of the head shake. *Gesture*, 2:147–182(36), 2002.

[10] M. Kipp, M. Neff, K. H. Kipp, and I. Albrecht. Towards natural gesture synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis. In *Proceedings of the 7th International Conference on Intelligent Virtual Agents*, pages 15–28, 2007.

[11] M. Knapp and J. Hall. *Nonverbal Communication in Human Interaction*. Harcourt Brace College Publishers, 4th edition, 1997.

[12] F. L. A. Knoppel, A. S. Tigelaar, D. O. Bos, T. Alofs, and Z. Ruttkay. Trackside DEIRA: a dynamic engaging intelligent reporter agent. In *AAMAS '08: Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems*, pages 112–119, Richland, SC, 2008. International Foundation for Autonomous Agents and Multiagent Systems.

[13] S. Kopp, B. Krenn, S. Marsella, A. Marshall, C. Pelachaud, H. Pirker, K. Thorisson, and H. Vilhjálmsson. Towards a common framework for multimodal generation in embodied conversation agents: a behavior markup language. In *Proceedings of 6th International Conference on Virtual Agents*, pages 205–217, Marina del Rey, CA, USA, 2006.

[14] J. Lee and S. Marsella. Nonverbal behavior generator for embodied conversational agents. In *Proceedings of the 6th International Conference on Intelligent Virtual Agents*, pages 243–255. Springer, 2006.

[15] J. Lee and S. Marsella. Learning a model of speaker head nods using gesture corpora. In *AAMAS '09: Proceedings of the 8th international joint conference on Autonomous agents and multiagent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2009.

[16] J. Lee, H. Prendinger, A. Neviarouskaya, and S. Marsella. Learning models of speaker head nods with affective information. In *ACII '09: Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction*. International Foundation for Autonomous Agents and Multiagent Systems, 2009.

[17] E. Z. McClave. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32:855–878(24), June 2000.

[18] A. Mignault and A. Chaudhuri. The many faces of a neutral face: Head tilt and perception of dominance and emotion. *Journal of Nonverbal Behavior*, 2(27):111–132, June 2003.

[19] L.-P. Morency, I. de Kok, and J. Gratch. Predicting listener backchannels: A probabilistic multimodal approach. In *Proceedings of the 8th International Conference on Intelligent Virtual Agents*, pages 176–190, 2008.

[20] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson. Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, 15:133–137(5), February 2004.

[21] A. Neviarouskaya, H. Prendinger, and M. Ishizuka. Textual affect sensing for sociable and expressive online communication. In *ACII '07: Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*, pages 218–229, Berlin, Heidelberg, 2007. Springer-Verlag.

[22] M. Strauss and M. Kipp. Eric: a generic rule-based framework for an affective embodied commentary agent. In L. Padgham, D. C. Parkes, J. Müller, and S. Parsons, editors, *AAMAS '08: Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems*, pages 97–104, Richland, SC, 2008. International Foundation for Autonomous Agents and Multiagent Systems.

[23] G. Tom, P. Pettersen, T. Lau, T. Burton, and J. Cook. The role of overt head movement in the formation of affect. *Basic and Applied Social Psychology*, 12(3):281–289, 1991.