Modeling Speaker Behavior: A Comparison of Two Approaches

Jina Lee and Stacy Marsella

University of Southern California,
Institute for Creative Technologies,
12015 Waterfront Drive, Playa Vista, CA 90094, USA
jinal@usc.edu, marsella@ict.usc.edu

Abstract. Virtual agents are autonomous software characters that support social interactions with human users. With the emergence of better graphical representation and control over the virtual agent's embodiment, communication through nonverbal behaviors has become an active research area. Researchers have taken different approaches to author the behaviors of virtual agents. In this work, we present our machine learning-based approach to model nonverbal behaviors, in which we explore several different learning techniques (HMM, CRF, LDCRF) to predict speaker's head nods and eyebrow movements. Quantitative measurements show that LDCRF yields the best learning rate for both head nod and eyebrow movements. An evaluation study was also conducted to compare the behaviors generated by the Machine Learning-based models described in this paper to a Literature-based model.

1 Introduction

Virtual agents are autonomous software characters that support social interactions with human users. One of the main goals in virtual agent research is to emulate how humans interact face-to-face. Virtual agents use natural speech and gestures to convey intentions, express emotions, and interact with human users much as humans use speech and gesture to interact with each other. Communication through a virtual agent's nonverbal behaviors has become an active research area of increasing importance, especially as better graphical representation and control over the agent's body has supported richer and subtler expression.

To realize this, researchers have taken different approaches to author the behaviors of virtual agents that are adaptable and appropriate to the context of the interaction. One of the foremost approaches in modeling nonverbal behaviors is the *Literature-based* approach, including the Nonverbal Behavior Generator (NVBG) [9] and the BEAT system [4]. This approach uses findings from nonverbal behavior research obtained through observation and interpretation of human interactions and builds computational models that operationalize those findings. In many of these studies, researchers analyze the recordings of human interactions and try to manually find regularities in various behaviors including head movements, posture shifts, or gaze movements. One of the major challenges with

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 161–174, 2012.

this approach lies in the fact that the full complexity of the mapping between various behaviors and communicative functions conveyed through the behaviors is not described in the research literature. There are many factors that potentially affect our nonverbal behaviors such as emotion, personality, gender, physiological state, or social context, and revealing the impacts of these factors and the interdependency among them is an extremely challenging task.

Recently, there have been growing efforts to use machine learning techniques as tools to learn patterns of behaviors [1,2,7,10,11,12,14]. In this *Machine Learning* approach, instead of manually trying to find associations between various factors and nonverbal behaviors, automated processes are used to find features that are strongly associated with particular behaviors. Then, one can use those features to train models that will predict the occurrences of the behavior. One advantage of this approach is that the learning is flexible and can be customized to find patterns in specific context. For example, to learn behavior patterns of different cultures, we may train separate models on each culture's data. Another advantage is that since the learning process is automated, we can process vastly larger amount of data in a given amount of time compared to manual analysis. However, obtaining good annotated data is often the greatest challenge in this approach.

The goal of the work described in this paper is two-fold. First, we extend our prior work [10] [12] of modeling speaker head nods using a machine learning approach to investigate whether we can improve the learning. Previously we built hidden Markov models to predict when speaker head nods occur, regardless of their magnitudes. Here we explore additional machine learning techniques and feature sets to predict not only the uniform head nods but also the different nod magnitudes and eyebrow movements. Secondly, we conduct an evaluation study to investigate how the different modeling approaches (Literature-based vs. Machine learning) compare with each other by asking human subjects to rate the perception of a virtual agent through its behaviors.

The following section describes the research on modeling nonverbal behaviors for virtual agents including our previous work using the two different approaches. We then describe the extension to the machine learning approach by exploring different learning techniques and features to learn patterns of uniform nods, nods with different magnitudes, and eyebrow movements. Finally, we present the evaluation study that shows how the behaviors from different modeling approaches are perceived by human users.

2 Related Work

Research on virtual agent has taken different approaches to realize nonverbal behaviors of the virtual agent. Our previous work on the Nonverbal Behavior Generator (NVBG) [9] employs a literature-based approach to generate behaviors according to the communicative functions. The system incorporates a set of nonverbal behavior rules that map from the agents' communicative intentions to various nonverbal behaviors. The communicative intentions are derived from a range of sources, from the agent's cognitive reasoning, dialog processing and emotional state, to linguistic features of the utterance text. For instance,

the *Intensification Rule* in NVBG is triggered when the surface text includes intensifying expressions, suggested by words like 'very' or 'quite,' which then generates lowered eyebrow movements and a head nod. However, multiple rules could apply to one text segment, leading to rule conflicts. To resolve those conflicts, rules were prioritized using the frequency counts of feature/behavior correspondence extracted from corpora of human nonverbal behavior.

Others have employed a machine learning approach by using corpora of nonverbal behavior more extensively and developing probabilistic models that find the behavior patterns from data directly. Below we list a number of different machine learning techniques used for modeling nonverbal behaviors, which we also use for the work described in this paper.

Hidden Markov model (HMM) [19] is a statistical model that has been widely used in problems with temporal dynamics such as speech recognition, handwriting recognition, and natural language problems including part-of-speech tagging. A number of work in gesture modeling is based on hidden Markov models. Busso et al. [2] used features of actual speech to synthesize emotional head motion patterns for avatars. Our previous work used various linguistic features to predict speaker head nods and investigates the impact of using affective information during learning [10,11,12].

Conditional Random Fields (CRF) [8] relax HMM's conditional independence assumption and can learn long-range dependencies between input features, making it more suitable for various real-world problems. Morency et al. [14] trained CRFs to predict listener's head movements by using various multi-modal features of the human speaker (e.g. prosody, spoken words, eye gaze) and exploring different feature encoding methods (e.g. binary, step function, ramp function). Sminchisescu et al. [20] also used CRFs to classify human motion activities such as walking, jumping, or running using 2D image features and 3D human joint angle features.

Latent-Dynamic Conditional Random Fields (LDCRF) [15] incorporate hidden state variables, which allows them to learn the substructure of gestures, however, while requiring more time and data to train the models. Morency et al. [15] trained a LDCRF for recognizing head and eye gestures using rotational velocity of the head or eye gaze at a specific time frame. They have also trained support vector machines (SVM), HMM, and CRF and showed that LDCRF models yielded the best performance on visual gesture recognition task.

3 Modeling Head Nods and Eyebrow Movements

In this section, we explore multiple learning techniques and feature sets for learning models of speaker head nods and eyebrow movements. We first present the gesture corpus and features used for training, then describe the probabilistic models we learned and the learning results.

3.1 Gesture Corpus

AMI Meeting Corpus [3] was used in this work, which includes annotations of speaker transcript, dialog acts of utterances and timings of people's head

movements in addition to other data (e.g. video data) not used in this work. The original corpus was also extended by our own annotations; first, we manually annotated the dynamics of the nods (small, medium, big) and eyebrow movements (inner brow raise, outer brow raise, brow lowerer). Secondly, we processed the speaker transcript through a text processor to obtain additional features. The following describes the features that were used to train the probabilistic models in detail.

Syntactic Features: Syntactic features include part of speech tags (18 total), phrase boundaries (sentence start, noun phrase start, verb phrase start), and key lexical entities (7 cases). Key lexical entities consists keywords that are shown to have strong correlations with head movements [13]. Some examples include 'yes' for affirmation and 'very' for intensified expressions.

Dialogue Acts: Dialogue acts describe the communicative functions of each utterance and are extracted from the AMI Meeting Corpus (15 total).

Paralinguistic Features: These features are also obtained from the gesture corpus and includes *gaps*, *disfluencies*, and *vocal sounds*. Gaps are speech gaps during speaking turns, disfluency markers are discontinuity or disfluencies while uttering, and vocal sounds are nonverbal sounds such as laughing, throat noises, or other nonverbal vocalizations.

Semantic Category: The speech transcription was processed through the Linguistic Inquiry and Word Count (LIWC) [18] to obtain the various semantic categories of each word. These categories include psychological construct categories (e.g., affect, cognition, biological processes), personal concern categories (e.g., work, home, leisure activities), paralinguistic dimensions (e.g. assents, fillers, nonfluencies), and punctuation categories (periods, commas, etc.). There are a total of 75 such categories.

Here we define the ***Basic Feature Set*** to include syntactic features, dialogue acts, and paralinguistic features, which are features that are obtained through a shallow parsing of the utterance. To study the impact of word semantics on learning the speaker behavior, we also define the ***Extended Feature Set*** to include the semantic categories in addition to the basic feature set.

3.2 Training Process

The HCRF Library [5] was used to train HMM, CRF and LDCRF models. For each learning technique, separate models were learned for different nods and eyebrow movements (e.g. separate CRF models for general, small, and medium nods). For HMMs and LDCRFs, the number of hidden states tried out are 2-6. Approximately 70% of the annotation data were used as training set and about 30% were used as test set, keeping the annotations of a particular person in either the training set or the test set. The training set was further split to set aside a validation set through a 3-fold cross validation. The training set, test set, and validation set were constructed with a sampling rate of 10Hz.

Table 1. Performance of the nod models

Model Type	*Feature Set*	*F-score*	*Precision*	*Recall*
GENERAL NODS				
HMM	Feature Basic	0.1421	0.0834	0.4788
HMM	Feature Extended	0.1357	0.0763	0.6101
CRF	Feature Basic	0.2575	0.2489	0.2667
CRF	Feature Extended	0.2525	0.2177	0.3004
LDCRF	Feature Basic	0.3002	0.2489	0.3781
LDCRF	Feature Extended	0.2665	0.2196	0.3391
SMALL NODS				
CRF	Feature Basic	0.1148	0.0745	0.2507
CRF	Feature Extended	0.0937	0.0517	0.4991
LDCRF	Feature Basic	0.1404	0.0996	0.2375
LDCRF	Feature Extended	0.1273	0.0750	0.4198
MEDIUM NODS				
CRF	Feature Basic	0.0654	0.0679	0.0631
CRF	Feature Extended	0.0368	0.0189	0.7109
LDCRF	Feature Basic	0.0479	0.0297	0.1248
LDCRF	Feature Extended	0.0423	0.0223	0.4135

3.3 Results

A number of models with different combinations of learning algorithms and feature sets were learned to predict the speaker head nod and eyebrow movements. First the results of predicting general head nods are described. We define *general head nods* as nods regardless of their magnitudes, thus combining all the nods in the original data. Next we present the results of learning nods with different magnitudes and various eyebrow movements.

General Head Nods

Given the differences between HMM, CRF, and LDCRF as described above, the underlying assumption was that CRF models will achieve better performances than HMMs, and LDCRF models will perform better than CRF models. The performance of the learned models were measured through F-score, precision, and recall (see the top entries of Table 1). The HMM model using the basic feature set yielded an F-score of 0.1421. As expected, the CRF models performed better than HMMs, and the LDCRF models performed better than the CRF models. This implies that learning the extrinsic dynamics between nods is important. Furthermore, the results of CRF and LDCRF models using basic or extended feature sets show that the LDCRF models (best F-score: 0.3002) perform better than the CRF models (best F-score: 0.2575), emphasizing the importance of learning the hidden substructure of nodding patterns. However, using the extended features did not seem to have a strong impact in all three learning algorithms.

Since the learning results of HMMs were noticeably lower than those of CRFs and LDCRFs, in the subsequent learning of nod magnitudes and eyebrow movements, we only report the results of CRFs and LDCRFs.

Head Nod Magnitudes

In addition to the general head nods, separate models for different dynamics of the nods (small, medium) were also learned. Among all the nod instances in the data, 53.4% were small nods, 40.5% were medium nods, and 6.1% were big nods. Overall, due to the reduced size of sample points, the performances of the models are not as high as those of the general nod models. For small nod models, the LDCRF model using basic feature set achieved the best F-score (0.1404) similar to the general nod model. The CRF model with basic feature set achieved the best performance for medium nods (F-score 0.0654) with a marginal improvement over the LDCRF model with basic feature set. Learning models for big nods failed due to lack of enough data.

Eyebrow Movements

4 different types of eyebrow models were learned: inner brow raise (AU1) models, outer brow raise (AU2) models, eyebrow raise models (combining AU1 and AU2) and brow lowerer (AU4) models. Results of the learned models are shown in

Table 2. Performance of the eyebrow models

Model	Type	F-score	Precision	Recall
INNER EYEBROW RAISE (AU1)				
CRF	Feature Basic	0.1871	0.1645	0.2170
CRF	Feature Extended	0.1661	0.1690	0.1633
LDCRF	Feature Basic	0.2749	0.1761	0.6265
LDCRF	Feature Extended	0.2066	0.1756	0.2509
OUTER EYEBROW RAISE (AU2)				
CRF	Feature Basic	0.1019	0.0562	0.5452
CRF	Feature Extended	0.1015	0.0566	0.4914
LDCRF	Feature Basic	0.1079	0.0695	0.2411
LDCRF	Feature Extended	0.0977	0.0568	0.3505
EYEBROW RAISE (AU1 or AU2)				
CRF	Feature Basic	0.3280	0.2155	0.6863
CRF	Feature Extended	0.3281	0.2109	0.7389
LDCRF	Feature Basic	0.3421	0.2438	0.5734
LDCRF	Feature Extended	0.3270	0.2345	0.5402
BROW LOWERER (AU4)				
CRF	Feature Basic	0.0133	0.0286	0.0087
CRF	Feature Extended	0.0874	0.0603	0.1585
LDCRF	Feature Basic	0.0770	0.0425	0.4120
LDCRF	Feature Extended	0.0733	0.0416	0.3053

Table 2. The inner brow raise (AU1) models yielded better results than the outer brow raise (AU2) or brow lowerer models (AU4). Given that the inner brow raises and the outer brow raises are hard to distinguish from facial expressions and that the two are strongly correlated, the two data were combined and an eyebrow raise model was learned. This eyebrow raise resulted in an improved performance with an F-score of 0.3421. Except for the brow lowerer model, in all the other cases the LDCRF models resulted in better performances than the CRF models, revealing the importance of learning the hidden substructure of eyebrow movements. Similar to the nod models, the extended feature set did not improve the learning over the basic feature set. The outer brow raise models and brow lowerer models had relatively poorer performances, due to lack of enough data samples in the corpus. The inner brow raise (AU1) movements consisted of 70.7% of all the eyebrow movements in the data, compared to 17.1% and 12.2% for outer brow raise (AU2) and eyebrow lowerer (AU4) movements, respectively.

3.4 Discussion

This section presented extensions to our previous machine learning approach on modeling speaker head nods [10]. Here the focus was on comparing different learning algorithms (HMM, CRF, LDCRF), exploring new features, and learning the dynamics of head nods and expanding the learning to eyebrow movements. As expected the LDCRF models tended to outperform other learning techniques. Specifically, considering models with higher F-scores (above 0.25), the LDCRF models had better results than the CRF models, implying the importance of learning the hidden sub-structure of the nods. HMMs had poorer performance than CRF or LDCRF models, suggesting that there may be long-range dependencies between the input features in modeling head movements and eyebrow movements. However, the extended feature set including additional semantic category did not improve the learning. Possible explanations for this could be that either the extended feature set was not very selective in learning these behaviors or that we need more data for the models to learn the relationships among different features.

4 Evaluation Study

This section presents an evaluation study on how the behaviors generated by different models are perceived by human users. Instead of comparing the behaviors from different probabilistic models, we take a broader view and address the issue of comparing different modeling approaches, namely the Literature-based approach and the Machine Learning approach. More specifically, we ask whether the different modeling approaches have an impact on the user's perception of the virtual agent. However, how to measure people's perception of the agent remains a challenge; simply asking subjects about naturalness, precision, and recall and asking them to make broad judgements about the behaviors, as done in an earlier study [11] raises questions. Recent work has moved to using instruments

based on social psychology research [1,6]. In line with this research, here we use a modification of measurements developed by Nass et al. on politeness [17] and Osgood et al. on semantic differential [17].

4.1 Study Design

Hypothesis

The main interest of this study is to investigate how the behaviors generated by the different models impact the perception of the virtual agent. Our previous study [11] comparing the speaker head nods from a Machine Learning-based model to a Literature-based model showed that the general nods from simpler HMMs were perceived to be higher in precision, recall, and more natural. Based on this, we hypothesize that the behaviors from the Machine Learning-based model will receive higher perceptual ratings than the behaviors from the Literature-based model.

Independent Variables

The independent variables are the different modeling approaches from which the behaviors are generated:

- *C1: Probabilistic models described in Section 3*
- *C2: Literature-based model (NVBG) [9]*

Stimuli

Each participant was assigned to one condition (i.e. between-subject design). As stimuli, video clips were generated of a virtual agent displaying head nods and eyebrow movements while speaking an utterance. 23 utterances were initially selected from the ICT's Virtual Human Toolkit [21] utterance set, in which a virtual agent answers questions about the concept of virtual humans, what the Virtual Human Toolkit is, and about himself. To generate the head nods and eyebrow movements for each conditions, there were four different models to choose from (CRF-Basic Feature Set, CRF-Extended Feature Set, LDCRF-Basic Feature Set, LDCRF-Extended Feature Set) for each behavior. Based on past observations of behavior generation methods, human users preferred virtual agents with more behaviors mainly because they made the agent look more alive and less robotic. This suggested using the model that generates the most behaviors, namely the model with the best recall rate among the four different choices. To validate this, a preliminary evaluation was conducted and confirmed that human users indeed preferred the behaviors generated from models with high recall rate[1].

[1] To validate the choice of using models with high recall rate, a preliminary evaluation study was conducted. Fifteen utterances were selected from the utterance set mentioned above and two versions of the videos were created displaying behaviors generated from 1) models with the highest F-score and 2) models with the highest recall rate. 12 participants were recruited and asked to choose the video they preferred in a forced-choice manner. In 13 videos sets out of 15, the video displaying behaviors from models with high recall rate was preferred.

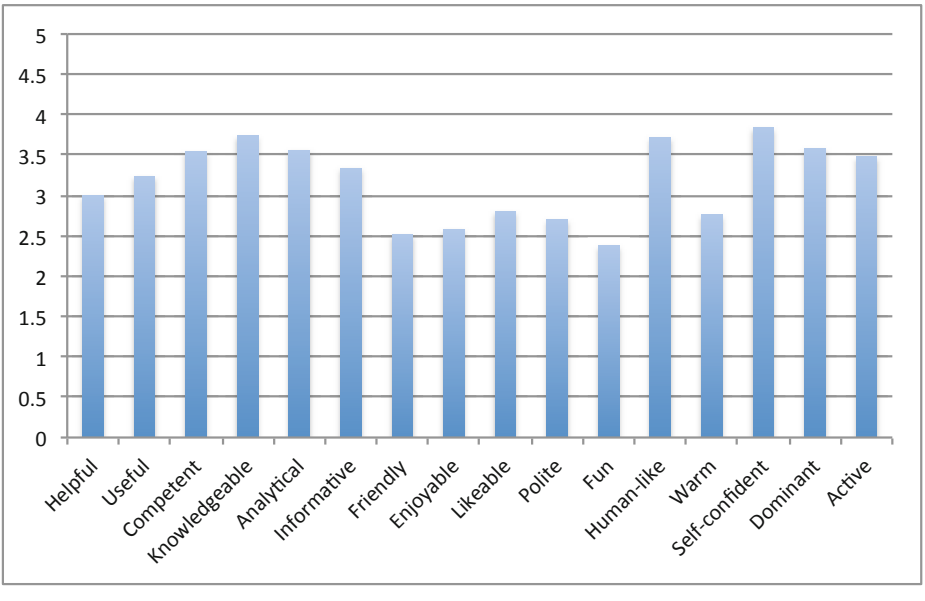


Fig. 1. Virtual agent Utah and the ratings on Utah's personality dimensions from a static image

Among the 23 utterances, 7 utterances with the greatest number of behavioral differences between the probabilistic model and NVBG were selected. In total, there were 14 video clips (7 utterances x 2 conditions) and each participant watched 7 video clips from the experimental condition they were assigned to.

Dependent Variables

To measure participants' perception of the agent through its behaviors, participants were asked to rate the agent on 16 personality dimensions (see Fig. 1) based on the studies of [6] and [1] using a 5-point Likert scale. These dimensions

are adopted from the study of Nass et al. [16] used in their politeness study and the semantic differential established by Osgood, Suci, and Tannenbaum [17].

Baseline

To measure the first impression of the virtual agent 'Utah' used in this study, a preliminary study was conducted. A separate set of participants were recruited online (30 males, 20 females, ages ranging from 18 to 65) and were asked to rate him on the 16 dimensions described above on a 5-point Likert scale after seeing a static image of Utah. Fig. 1 shows the mean values of the ratings for each personality dimension. Utah was rated high on *competent, knowledgeable, analytical, informative, human-like, self-confident, dominant*, and *active* but low on *friendly, enjoyable, likeable, polite, fun*, and *warm.* The evaluation results presented in the next section are based on the differences between these means of Utah's initial impression and the ratings of the human subjects to measure the effects of only the modeling conditions and not the appearance of Utah.

Procedure

90 participants were recruited online with 46 males and 44 females and ages ranging from 18 to 65. Participants first filled out a demographic questionnaire asking for their age, gender, education level, ethnicity, and occupation. They were assigned to one modeling condition and watched 7 video clips of the agent speaking a sentence while making head nods and eyebrow movements. Each video clip lasted about 10 seconds. After watching each video, participants were asked to rate the agent on the 16 personality dimensions using a 5-point Likert scale. The order of the video clips and the 16 dimension ratings were randomized for each participant.

4.2 Results

First the means of the personality dimensions obtained in the preliminary study (using Utah's static image) were subtracted from the participants' ratings to provide a more accurate measurement of the perception of the agent due to the behaviors generated by different models.

To measure the reliability between the dimensions, the 16 dimensions were grouped into several factors by conducting a factor analysis. Three factors were extracted, explaining 68.75% of the total variance. Calculating Cronbach's alpha showed that the alpha values for all three factors were above 0.7, justifying combining the dimensions in the same group as a single value (see Table 3). These factors were labeled as Competence, Likeability, and Power.

Independent-samples T-tests were conducted to study the main effect of the modeling approach. The mean values and standard deviations are shown and plotted in Fig. 2. Between the two modeling approaches, there was a significant difference in Power, where the Literature-based model was rated significantly

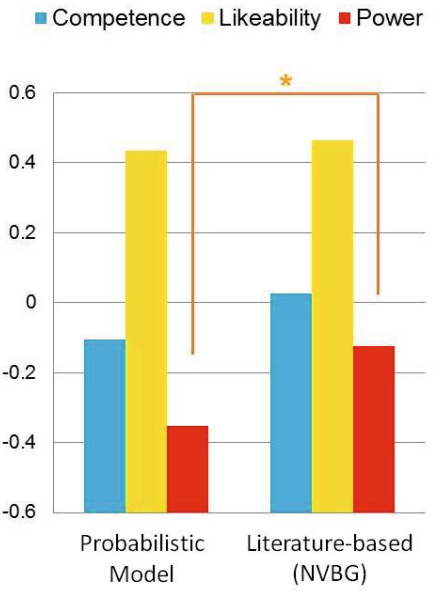
Table 3. Factor analysis for the perception of Utah's personality (principal component analysis with varimax rotation)

Item	Competence	Likeability	Power
Helpful	.805		
Useful	.804		
Competent	.774		
Knowledgeable	.811		
Analytical	.705		
Informative	.764		
Friendly		.826	
Enjoyable		.843	
Likeable		.851	
Polite		.670	
Fun		.765	
Human-like		.502	
Warm		.710	
Self-confident			.764
Dominant			.863
Active			.592
Cronbachs α	*.901*	*.906*	*.790*

higher (t(675.406) = -3.442, p<.01). However, there was no significant difference in Competence and Likeability. Therefore, the hypothesis stating that the Machine learning-based model will outperform the Literature-based model was *not* supported. In fact there was a trend that the Literature-based was rated higher than the Machine learning-based model in Competence, Likability, and Power.

4.3 Discussion

The evaluation result contradicts that of the earlier study [11] that contrasted the Literature-based model with the simpler HMM model, with the HMM being judged superior. There are several possible explanations. The earlier HMM study was attempting to learn just the head nods regardless of their magnitudes, whereas the current study involves head nods of different magnitudes and eyebrow movements as well. These finer behavioral distinctions perhaps made it harder for the models to learn the behavior patterns since there are fewer consistent behaviors across different subjects. In addition, showing more behaviors adds complexity to the evaluation task since there are more factors to consider when judging the perception of the virtual agent. The hypothesis was also setting a tough criteria for the Machine learning-based model. First, it required the Machine learning-based model to be rated higher in all three factors. Perhaps what the hypothesis should have focused on was the Likeability factor, which groups dimensions such as *Friendly, Human-like*, and *Warm*. Comparing the ratings for Likeability between the Machine learning model and the Literature-based model, the two models received similar ratings with a marginal difference. Second, the mappings and behaviors of the Literature-based model (NVBG) were



	Competence	Likeability	Power
C1 Probabilistic Model	-.104 (.913)	.434 (.797)	-.353 (.808)
C2 Literature based Model	.027 (.851)	.465 (.878)	-.125 (.924)

Fig. 2. Mean values of the agent perception factors. The means reflect differences from the agent's initial impression, shown in Fig. 1. Significant differences are identified by * ($P<.05$) and the exact values shown in the table (standard deviations shown in parenthesis).

distilled from years of social psychology research. Furthermore, it has been constantly modified and refined over the years as NVBG has been incorporated into numerous virtual agent projects, altogether setting a high bar for the machine learning model to beat.

5 Conclusion

In this paper we presented the work on learning probabilistic models to predict speaker head nods and eyebrow movements. We explored different learning algorithms (HMM, CRF, LDCRF) and feature sets to learn when speaker nods occur, as well as to learn the dynamics of head nods and the eyebrow movements. Consistent with our expectations, quantitative results (e.g. F-score) show that the LDCRF models had the best results, implying the importance of learning the dynamics between different gesture classes and the hidden sub-structure of the gestures. However, the extended feature set including additional semantic categories did not improve the learning, perhaps due to the fact that it requires more data to learn the impacts of the additional features.

The evaluation study conducted with human subjects focused on investigating how the behaviors generated by the different models affect the perception of the agent. Contrary to our expectation, the Machine learning-based model did not receive higher ratings than the Literature-based model; the complexity of the behavior sets in the video and the tough criteria to support the hypothesis may explain this result. This demands a follow-up study.

This work could be extended in several ways. Similar probabilistic approaches could be taken to learn patterns of additional behaviors or mappings of different communicative functions. For example, we may customize the learning by training models on data from specific groups of people that convey their status, individual traits, or cultural background. The evaluation study could also be improved by letting the human users *interact* with the virtual agents rather than showing videos of them. In addition, a more comprehensive evaluation is necessary to study the implications of each type of behavior generated under different conditions and what the users infer from each of those behaviors.

References

1. Bergmann, K., Kopp, S., Eyssel, F.: Individualized Gesturing Outperforms Average Gesturing – Evaluating Gesture Production in Virtual Humans. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS, vol. 6356, pp. 104–117. Springer, Heidelberg (2010)
2. Busso, C., Deng, Z., Grimm, M., Neumann, U., Narayanan, S.: Rigid head motion in expressive speech animation: Analysis and synthesis. IEEE Transactions on Audio, Speech and Language Processing 15(3), 1075–1086 (2007)
3. Carletta, J.: Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. Language Resources and Evaluation Journal 41(2), 181–190 (2007)
4. Cassell, J., Vilhjálmsson, H.H., Bickmore, T.: BEAT: the behavior expression animation toolkit. In: SIGGRAPH 2001: Proc. of the 28th Annual Conf. on Computer Graphics and Interactive Techniques, pp. 477–486 (2001)
5. HCRF library (including CRF and LDCRF) (2012), `http://sourceforge.net/projects/hcrf/`
6. Hoffmann, L., Krämer, N.C., Lam-chi, A., Kopp, S.: Media Equation Revisited: Do Users Show Polite Reactions towards an Embodied Agent? In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 159–165. Springer, Heidelberg (2009)
7. Kipp, M., Neff, M., Kipp, K.H., Albrecht, I.: Towards Natural Gesture Synthesis: Evaluating Gesture Units in a Data-Driven Approach to Gesture Synthesis. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 15–28. Springer, Heidelberg (2007)
8. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of the Eighteenth Int. Conf. on Machine Learning, pp. 282–289 (2001)
9. Lee, J., Marsella, S.C.: Nonverbal Behavior Generator for Embodied Conversational Agents. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 243–255. Springer, Heidelberg (2006)

10. Lee, J., Marsella, S.: Learning a model of speaker head nods using gesture corpora. In: Proc. of the 8th Int. Joint Conf. on Autonomous Agents and Multiagent Systems (2009)
11. Lee, J., Marsella, S.C.: Predicting speaker head nods and the effects of affective information. IEEE Transactions on Multimedia 12(6), 552–562 (2010)
12. Lee, J., Neviarouskaya, A., Prendinger, H., Marsella, S.: Learning models of speaker head nods with affective information. In: Proc. of the 3rd Int. Conf. on Affective Computing and Intelligent Interaction (2009)
13. McClave, E.Z.: Linguistic functions of head movements in the context of speech. Journal of Pragmatics 32, 855–878 (2000)
14. Morency, L.-P., de Kok, I., Gratch, J.: Predicting Listener Backchannels: A Probabilistic Multimodal Approach. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 176–190. Springer, Heidelberg (2008)
15. Morency, L.P., Quattoni, A., Darrell, T.: Latent-dynamic discriminative models for continuous gesture recognition. In: IEEE Conf. on Computer Vision and Pattern Recognition (2007)
16. Nass, C., Moon, Y., Carney, P.: Are People Polite to Computers? Responses to Computer-Based Interviewing Systems. Journal of Applied Social Psychology 29(5), 1093–1109 (1999)
17. Osgood, C.E., Suci, G.J., Tannenbaum, P.H.: The measurement of meaning, p. 197. University of Illinois Press (1957)
18. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic Inquiry and Word Count: LIWC 2001. Word Journal of the International Linguistic Association (2001)
19. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
20. Sminchisescu, C., Kanaujia, A., Li, Z., Metaxas, D.: Conditional models for contextual human motion recognition. In: Int. Conf. on Computer Vision, pp. 1808–1815 (2005)
21. ICT Virtual Human Toolkit (2012), `http://vhtoolkit.ict.usc.edu`