Learning a Model of Speaker Head Nods using Gesture Corpora

Jina Lee
Institute for Creative Technologies
University of Southern California
13274 Fiji Way, Marina del Rey, CA 90292 USA
jlee@ict.usc.edu

Stacy Marsella
Institute for Creative Technologies
University of Southern California
13274 Fiji Way, Marina del Rey, CA 90292 USA
marsella@ict.usc.edu

ABSTRACT

During face-to-face conversation, the speaker's head is continually in motion. These movements serve a variety of important communicative functions. Our goal is to develop a model of the speaker's head movements that can be used to generate head movements for virtual agents based on a gesture annotation corpora. In this paper, we focus on the first step of the head movement generation process: predicting when the speaker should use head nods. We describe our machine-learning approach that creates a head nod model from annotated corpora of face-to-face human interaction, relying on the linguistic features of the surface text. We also describe the feature selection process, training process, and the evaluation of the learned model with test data in detail. The result shows that the model is able to predict head nods with high precision and recall.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; I.2.11 [**Distributed Artificial Intelligence**]: Intelligent agents

General Terms

Design, Human Factors

Keywords

Virtual Agents, Embodied Conversational Agents, Nonverbal Behaviors, Head Nods, Machine Learning

1. INTRODUCTION

During face-to-face conversation, the head is constantly in motion, especially during speaking turns [12]. These movements are not random; research has identified a number of important functions served by head movements [24] [17] [13] [14]. Head movements provide a range of information in addition to the verbal channel. We may nod to show our agreement with what the other is saying, shake our heads to express disbelief, or tilt the head upwards along with gaze aversion when pondering something. In addition to serving these explicit functions, head movements may also influence the observer in more subtle ways. For example, overt head movements is found to be instrumental in the formation of an observer's affective response to the speaker [32]. Additionally, the various head movements we make during conversation make the interaction look more natural.

Consistent with the important role that head movements play in human-human interaction, virtual agent systems have incorporated head movements to realize a variety of functions [1] [4] [5] [10] [20] [21] [30]. The incorporation of appropriate head movements in a virtual agent has been shown to have positive effects during human-agent interaction [27]. The goal of our work is to build a domain-independent model of speaker's head movements that can be used to generate head movements for virtual agents. To use the model for interactive virtual agents, we design it to work in real-time and to be flexible enough to be used in different virtual agent systems.

Often virtual humans use hand-crafted models to generate head movements. For instance, in our previous work we developed the Nonverbal Behavior Generator (NVBG) [21], which is a rule-based system that analyzes the information on the agent's cognitive processing, such as its internal goals and emotional state, but also analyzes the syntactic and semantic structure of the surface text to generate a range of nonverbal behaviors. To specify which nonverbal behaviors should be generated at each given context, the knowledge from the psychological literature and analysis of human nonverbal behavior corpora are used to identify the salient factors most likely to be associated with certain nonverbal behaviors.

As with a number of systems [1] [4] [5] [20] that generate nonverbal behaviors for virtual humans, the NVBG work starts with specific factors that would cause various gestures to be displayed. Although the knowledge encoded in the NVBG rules has been reused and demonstrated to be effective across a range of applications [31] [33] [18] [15], 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.

Cite as: Learning a Model of Speaker Head Nods using Gesture Corpora, Jina Lee, Stacy Marsella, *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009),* Decker, Sichman, Sierra and Castelfranchi (eds.), May, 10–15, 2009, Budapest, Hungary, pp. 289–296

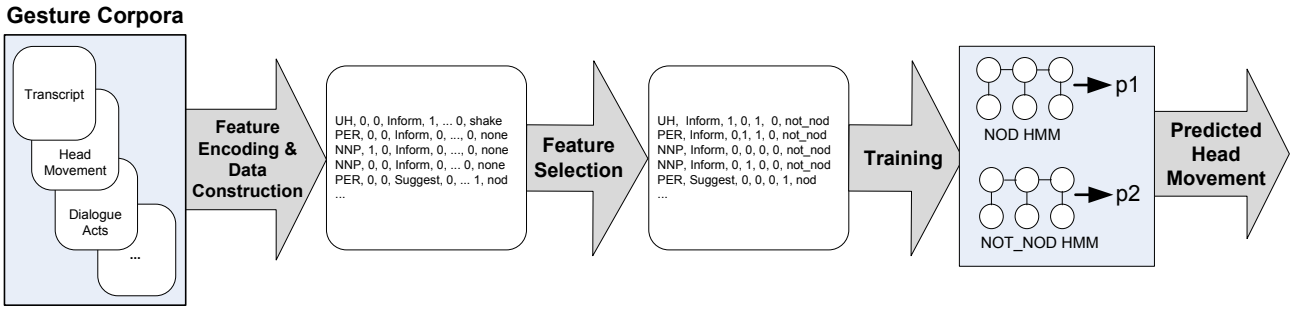

Figure 1: Overview of the head nod prediction framework. The information in the gesture corpus is encoded and aligned to construct the data set. The feature selection process chooses a subset of the features that are most correlated with head nods. Using these features, probabilistic sequential models are trained and utilized to predict whether or not a head nod should occur.

To complement the limitations of our previous rule-based approach, we present a data-driven, automated approach to generate speaker nonverbal behaviors, which we demonstrate and evaluate. Specifically, the approach uses a machine learning technique (i.e. learning a hidden Markov model [29]) to create a head nod model from annotated corpora of face-to-face human interaction. Because our goal is a flexible system that can be used in different virtual agent systems with various approaches to natural language generation, we restrict the features used in the machine learning to those available across different systems. Specifically, we explore in this paper the use of features available through shallow parsing and interpretation of the surface text and leave for future work the exploration of deeper features.

There are several advantages with this machine learning approach. First of all, the process is automated. Having a good understanding of the phenomena is still important, however with this 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. What becomes more important is the process of choosing the right features to train the model. In this work, we focus on the linguistic features of the surface text when learning the model rather than, for example, visual feedback. In other words, we would like to use for training the same information from the natural language generator that will be available when the learned model is incorporated into a virtual human. Another advantage of this approach is that it is flexible and can be customized to learn for a specific context. For example, if we want to learn the head nod patterns of different cultures, we may train each model 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, as done in [19]. The advantages of machine-learning approach makes it a strong alternative to rule-based approach or a substantial enhancement when both are used.

In this paper, we describe our approach for learning to predict the speaker's head nods from gesture corpora. Once the patterns of when people nod are learned, we can use the model to automatically encode a new sample with the best features used for learning to generate the most likely head movement for virtual agents. Although the focus in this paper is on the initial steps of learning and evaluating the model, the model could in turn be incorporated into a larger system like NVBG.

The following section describes the research on head movements, previous work on modeling head movements for virtual agents, and the diverse approaches each system employs. We then describe our approach in detail, including the data construction process, feature selection process, training process, as well as the evaluation of the learned model with test data. Figure 1 depicts the overview of the procedures to learn the model. The results show that the model is able to predict head nods with high precision and recall. 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 [13] [14] [17] [24]. Heylen [14] 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 [17] 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 [24], 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 speakers. Speakers also predictably change the head position when discussing alternatives or items in a list.

Following the studies on nonverbal behaviors, many virtual

agents model these behaviors. Some generate the behaviors according to the 'conversation phenomena' or discourse structure. REA's [4] verbal/nonverbal behaviors are designed in terms of conversational functions. Rea employs head nods for sending feedbacks and head toss for signalling 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. Breitfuss et al. [1] developed 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.

Other virtual agents focus on generating expressive behaviors according to the agent's emotional state. Mancini et al. [23] show how complex emotion could be displayed through head movements driven by music expressivity. They use acoustic cues and emotions to show how musical expressivity could be transformed to behavioral expressivity. Deira [20] is a reporter agent that generates basic head movements (including facial expressions) at fixed intervals and but also produces more pronounced movements as the agent's excitement rises during the report. Similarly, ERIC [30] is a commentary agent that shows 'idle' gestures when no other gestures are requested, but generates various nonverbal behaviors according to its emotional state.

As mentioned in the previous section, Nonverbal Behavior Generator [21] 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 our own study of corpora of human nonverbal behaviors. Recently, there have been growing efforts to use corpora of nonverbal behavior more extensively. Morency et al. [26] creates a model that predicts listener's backchannel head nods using the speaker's multi-modal features (e.g. prosody, spoken words, eye gaze). Similarly, [34] [28] [6] and [22] also uses prosodic features to predict listener's backchannel head nod. Busso et al. [2] use audiovisual signals to synthesize emotional head motion patterns. They use prosodic features and facial expressions recorded from human speakers to build hidden Markov models for each emotional categories and use those models to synthesize head motions. The generated head motions are illustrated through an animated face. Their evaluation shows that head motion modifies emotional perception of facial animation especially in valence and activation domain. Kipp et al. [19] 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.

Foster and Oberlander [10] also present a corpus-based generation of head and eyebrow motion for virtual agent. They recorded and annotated a corpus of facial expressions and head movements and used the data to synthesize facial displays on RUTH [9]. Their approach for generating the behaviors is similar to ours presented in this paper, however there are several differences. First of all, although their approach is data-driven, they do not use machine learning techniques to construct the model. Instead, they count the frequencies of behaviors observed for the same context (i.e. feature combination) and either choose the behavior that was most frequently observed or make a weighted choice among all the different behaviors observed. Secondly, the features they use are based in part on specific domain and language tools. The utterances in the corpus are about bathroom tile design, and one of the features they use is the user-preference evaluation of objects being presented (e.g. which tile shapes or designers each user prefers). Another feature used is the pitch accent information provided by their COMIC text planner [11]. The use of features tied to specific user information or language tool may limit the portability of their work to other virtual agent systems. For our work, we emphasize on the generality of the model and create a head nod model that is domain-independent and re-usable in other systems. To that end, we concentrate on features that are easily obtainable across systems using various language tools.

In the works described above, head nods for embodied agents are either generated to realize certain communicative functions the agent plans to deliver or by learning the patterns from real human data. In the first case, head nods occur to greet, emphasize certain points, or to express the agent's emotion. However, this approach may lead the agent to look rigid or unnatural when only a few gestures are used to deliver the communicative function that may span over several utterances. To avoid the agent from looking too robotic, many virtual agents add in random head movements, which do not serve any particular functions other than to make the agent look alive. On the other hand, behaviors generated by data-driven approach may be more natural looking. However, majority of the systems that implement this approach model the listener's backchannel movements, not the speaker's movements, or if not, use features produced by specific tools or assume natural speech input (that has already embedded prosodic information relevant to head nods).

As mentioned above, we want to model the speaker's head movements and use the learned model to generate head nods in real time for virtual agents. For this reason, we focus on features that are readily available at the time head movements are generated. In addition, we plan to make the model portable to other systems by using features such as part of speech tags that are easily obtainable even when using different language tools. Using additional information such as pitch accents or facial expressions may greatly improve the learning, but the natural language generator may not generate those information or they may not be available at the time the model is used to generate nods. For this work, we emphasize on the portability and generality of the model and implement a minimalist approach. In the following section, we show that even with shallow model of the surface text, we can learn the model of speaker's head nods with high values of performance measures.

3. PREDICTING SPEAKER HEAD NODS

In this section, we describe our machine learning approach for learning the speaker head nods. First we describe the gesture corpus we used, followed by the feature selection process. Finally, we give a detailed description on how we trained the model and the results of the trained model.

3.1 Gesture Corpus

Figure 2: Snapshot of the meeting setting used for AMI meeting corpus [3].

The AMI Meeting Project is a European-funded multi disciplinary consortium formed to promote the research of group interaction [3]. The AMI Meeting Corpus is a set of multimodal meeting records, which includes 100 meeting hours. Each meeting consists of three or four participants placed in a meeting-room setting with microphones, a slide projector, electronic whiteboards, and individualized and room-view cameras. Figure 2 shows the meeting setting from which the corpus was created. There are two types of meetings in the corpus: scenario meetings and non-scenario meetings. In the scenario meetings, participants play the roles of employees in an electronics company and discuss the development of a new television remote control. Each participant plays a specific role (e.g. project manager, marketing expert, user interface designer, etc.) and are provided information from the scenario controller about when to start and finish the meetings, what to prepare for the meetings, etc. There are no scripts given to the participants. In the non-scenario meetings, participants are colleagues from the same area and have discussions on their research topics (e.g. speech research colleagues discussing posterior probability methods). Again, no script is given to the participants.

The corpus includes annotations of meeting context such as participant IDs and topic segmentations as well as annotations on each participant's transcript and movements. Annotations of each meeting are structured in an XML format and are cross-referenced through meeting IDs, participant IDs, and time reference. The following lists some of the annotations with brief descriptions (not a complete list).

- Dialogue Acts: Speaker intentions such as information exchange, social acts, and non-intentional acts.
- Topic Segmentation: A shallow hierarchical decomposition into subtopics (e.g. opening of meeting, chitchat).
- Named Entities: Codes for entities (people, locations, artifacts, etc.) and time durations (dates, times, durations).
- Head Gestures: Head movements of each participant.
- Hand Gestures: Hand movements of each participant.
- Movement: Abstract description of participant's movements (e.g. sit, take_notes, other).
- Focus of Attention: Participant's head orientation and eye gaze.
- Words: Transcript of words spoken by each participant.

1	ES2003a.A	ES2003a.B		
2	ES2003b.A	ES2003b.B	ES2003b.C	ES2003b.D
3	ES2008a.A	ES2008a.B	ES2008a.C	ES2008a.D
4	ES2008b.A	ES2008b.B	ES2008b.C	ES2008b.D
5	ES2008c.A	ES2008c.B	ES2008c.C	
6	ES2008d.A	ES2008d.B	ES2008d.C	ES2008d.D
7	ES2009a.A	ES2009a.B	ES2009a.C	ES2009a.D
8	ES2009b.A	ES2009b.B	ES2009b.C	ES2009b.D
9	ES2009c.A	ES2009c.B	ES2009c.C	ES2009c.D
10	ES2009d.A	ES2009d.B		
11	IS1000a.A	IS1000aB	IS1000a.C	IS1000a.D
12	IS1000b.A	IS1000b.B	IS1000b.C	IS1000b.D
13	IS1001a.A	IS1001a.B	IS1001a.C	IS1001a.D
14	IS1001b.A	IS1001b.B	IS1001b.C	IS1001b.D
15	IS1001c.A	IS1001c.B	IS1001c.C	
16	IS1001d.A	IS1001d.B	IS1001d.C	IS1001d.D
17	IS1002b.A	IS1002b.B	IS1002b.C	IS1002b.D

Table 1: List of meeting annotations [3] used for learning. Recordings of 17 meetings were used, which adds up to be around eight hours of annotation.

Assess	Elicit-Inform
Backchannel	Elicit-Offer-Or-Suggestion
Inform	Elicit-Assessment
Fragment	Elicit-Comment-Understanding
Offer	Comment-About-Understanding
Be-Positive	Be-Negative
Stall	Suggest
Other	

Table 2: Types of dialog act labels used in the corpus.

3.2 Data Alignment and Feature Selection

For this work, we used the recordings of 17 meetings, each consisted of three to four participants, which adds up to be around eight hours of meeting annotation. The meetings used for learning are listed in Table 1.

One of the main features of the earlier NVBG work was its robustness, the ability to generate behaviors even when all that was available was the surface text and a minimal set of information about the virtual human's internal state. Here, we take a similar approach; specifically, a shallow parsing is performed to analyze the syntactic and semantic structure of the surface string to predict head nods. Among all the annotations included in the corpus, we used the transcript of each speaker, the dialog acts of each utterance, and the type of head movements observed while the utterance was spoken. Table 2 lists the different types of dialogue acts used in the corpus. The head types annotated in the corpus are: nod, shake, nodshake, other, and none. Snapshots of the head movements are shown in Figure 3. We also obtained the part of speech tags and phrase boundaries (e.g. start/end of verb phrases and noun phrases) by sending the utterances through a natural language parser (Charniak Parser [8]). In

Figure 3: Snapshots of head movements in AMI corpus [3]. From the top: *nod*, *shake*, *nodshake*, and *other* head movements.

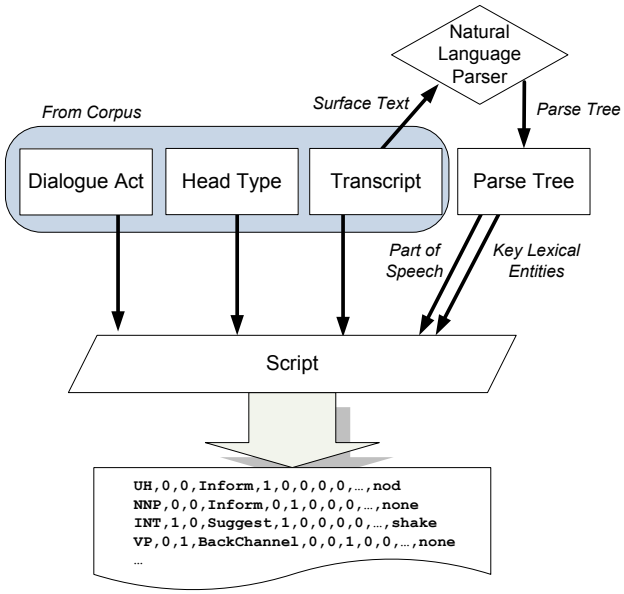

Figure 4: Data Construction Process. From the gesture corpus, speaker transcript, dialog act, and head types are extracted. The transcript is sent to the natural language parser to extract the part of speech tags and phrase boundaries. A script automatically cross-references each file to construct the data set. This data set is encoded and transformed into trigrams before being used to train the HMMs.

addition, we also combined the features from the nonverbal behavior rules used in NVBG; specifically, we looked for keywords that are shown to be associated with head nods in our prior work. We call those keywords *key lexical entities*. Figure 4 illustrates the data construction process.

From the 17 meeting recordings we used, we collected 10,000 sentences and wrote a script to cross-reference the corresponding annotation files and aligned the features at the word level. In other words, we aligned each word with the following:

- Part of speech tag (29 cases)
- Dialog act
- Phrase Boundaries: sentence start/end, noun phrase start, verb phrase start
- Key lexical Entities (whether or not the word triggers NVB rules associated with head nods)

For the particular kind of model we are training (i.e. hidden Markov models), adding another feature means we need more data samples to learn the combinations of all the features and how they affect the outcome we are trying to classify. With a limited number of data samples, we want to keep the number of features low by eliminating uncorrelated features (i.e. features that do not affect head nods). Therefore, we reduced the number of features by counting the frequency of head nods that occurred with each feature and selected a subset of them. Table 3 lists the frequency counts of these features (out of 2590 words with nods). Head nods

Part of Speech	
Interjection	427
Proper Noun	300
Conjunction	239
Adverb	238

Dialog Act	
Inform	910
BackChannel	387
Suggest	265

Phrase Boundaries	
sentence_start	2268
np_start	493
vp_start	391

NVBG Rule	
key_lexical_entities	594

Table 3: Features that most frequently co-occurred with head nods from the gesture corpus (Out of 2590 words that co-occurred with nods). The frequency counts are independent from each other.

Part of Speech	Conjunction, Proper Noun, Adverb, Interjection, Remainder
Dialog Act	BackChannel, Inform, Suggest, Remainder
Sentence Start	y, n
Noun Phrase Start	y, n
Verb Phrase Start	y, n
Key Lexical Entities	y, n

Table 4: Features selected for training. The features were selected based on the results of Table 3. The label 'Remainder' includes everything not falling under other categories.

occurred more frequently at the beginnings of utterances and noun/verb phrases than at the end of each. From part of speech tags, *Interjection* was most correlated with head nods, followed by *Proper Nouns*, *Conjunctions*, and *Adverbs*. Dialog Act *Inform* most frequently co-occurred with nods along with *BackChannel* and *Suggest*. There was also a substantial number of nods occurring with the *Key Lexical Entities* (keywords), confirming the validity of NVBG rules associated with head nods. Based on the results described above, the final features were selected for training. Table 4 lists the final features.

3.3 Training Process

To learn the head nod model, hidden Markov models (HMM) [29] were trained. HMM is a statistical model that is widely used for learning patterns where a sequence of observations is given. Some of the applications where HMM have been successfully used are gesture recognition, speech recognition, and part-of-speech tagging [35] [16] [7]. For this work, the input is a sequence of feature combinations representing each word. The sequential property of this problem led us to use HMMs to predict head nods.

After aligning each word of the utterances with the selected features, we put together a sequence of three words to form a set of trigrams, which would be used as our data set. For each trigram, the head type was determined by the majority vote method. For example, if more than two out of three words co-occurred with a nod, the trigram was classified as a nod instance, and the same applied for other head movement types. To determine whether a trigram should be classified as a nod, we trained two HMMs: a 'NOD HMM' and a 'NOT_NOD HMM,' which includes trigrams with head types other than a nod. Since the output of an HMM is a probability that a sample is labeled with a particular classification, we feed the same trigram into both models and compare the probabilities to determine its classification.

To train a 'NOD HMM,' we collected all the positive instances of 'nod' trigrams from the entire set of trigrams. Then, we left out 20% of the 'nod' trigrams as a test set, which is used in the final evaluation step, and used the remaining 80% of the data for training. To determine the parameter setting of HMM (i.e. the number of hidden states) that produces the best result, we performed a 10-fold cross-over validation for each parameter setting. That is, we split the remaining 80% of the data into 10 parts and used one part as a validation set and 9 parts as a training set. After training the model, we obtained measurements of the model. We repeated this process 10 times and obtained an average measurement for the given number of hidden states. By comparing the average measurements, we then determined the best number of hidden states. After this, we combined all 10 parts and trained the final 'NOD HMM' with the chosen number of hidden states. Similarly, we collected the positive instances of 'NOT_NOD' trigrams (i.e. trigrams with head movements other than nod) and repeated the above steps to train a final 'NOT_NOD HMM.' Finally, we ran the test set (20% of the entire data left out) through the 'NOD HMM' and 'NOT_NOD HMM' and classified each sample to have the head movement of whichever model produced a higher probability.

Measurement	Equation	Value
Accuracy	(tp+tn) / (tp+fp+tn+fn)	.8528
Precision	tp / (tp+fp)	.8249
Recall	tp / (tp+fn)	.8957
F-measure	2*precision*recall / (precision+recall)	.8588

Table 5: Measurements for the performance of the learned model.

3.4 Results and Discussion

To measure the performance of our learned model, we computed the accuracy, precision, recall, and F-measure of the learned model. Accuracy is the ratio of samples that were correctly classified. Precision is the ratio between the number of actual nods in the data and the number of nods predicted by the learned model. Recall is the ratio between the number of nods predicted by the learned model and the number of nods in the actual data. For the F-measure, we gave the recall and precision the same weight (F_1). Table 5 summarizes the results with the equations used for computing the measurements. The results show that the model can predict head nods with high precision, recall, and accuracy rate with only a shallow model of the surface text (i.e. only using the syntactic/semantic structure of the utterance and the dialog act).

In addition to the main results presented in table 5, a second preliminary experiment was conducted to assess which features were more important in the model. We took out one feature at a time and trained the HMMs with the rest of the

	Precision	Recall	F-score
Proper Noun	0	0	0
Adverb	0.0009	0.0061	0.0034
Verb Phrase Start	0.0613	0.0123	0.0382
Noun Phrase Start	0.066	0.0061	0.0375
Suggest	0.0695	-0.0123	0.0301
Interjection	0.0757	-0.0061	0.0363

Table 6: Changes in Precision, Recall, F-score rates of selective features when each was taken out from learning. The changes are computed from the results in Table 5.

features. For each case, we computed the accuracy, precision, recall and F-score and compared them to the previous values by computing the differences in each measurements.

For many features, removing them had small trade-offs in precision and recall rates. However, some feature extractions had more notable impact. We show these in Table 6. Specifically, *Proper Noun* and *Adverb* did not affect the learning at all or very marginally when taken out, where as *Verb Phrase Start*, *Noun Phrase Start*, *Suggest*, and *Interjection* resulted in a larger change in both precision and f-score values when taken out from learning. Interestingly, in the case of *Verb Phrase Start* and *Noun Phrase Start*, in NVBG head nods were inserted in those places to make the agent look more life-like. This second experiment suggests two future directions in our work. It raises a need for a more sophisticated automatic feature selection process such as the method used by Morency et al. [26], which can investigate the correlations of the features and head nods more thoroughly than a simple frequency count. Additionally, further evaluation with human subjects is needed. For example, it may be that the behavior looks more natural if we include those *Noun Phrase Start* and *Verb Phrase Start* features even though the F-score drops.

4. CONCLUSIONS AND FUTURE DIRECTION

In this paper we presented an approach to learning a probabilistic model to predict head nods using a gesture corpus. As mentioned above, our goal is to use the model to generate head nods for virtual agents. In this paper, we focused on using the linguistic features of the surface text, including the syntactic/semantic structure of the utterance and other information that may be provided by the virtual agent's natural language generator. We trained hidden Markov models to predict head nods. The results show that the learned models predict head nods with high values of precision, recall and F-scores. A follow-up assessment explored what features had the most impact on head nod prediction.

This work shows that human head nods could be predicted with high performance measures using machine learning approach even without a rich markup of surface text. Compared to knowledge-intensive approach where the rule-author needs to manually construct rules that generate head nods, this approach does not require a complete knowledge of the correlations of the factors that may affect head nods. Instead, the author may concentrate on selecting the right features used for machine learning, which in our case was guided by the research on head movements.

This work could be extended in several ways. Currently we are working on detecting the emotional state from each utterance and adding this into the feature set to investigate whether emotional data improves the learning. Further analysis of the linguistic structure may also be performed using additional language tools to extract features such as emphasis points and contrast points. In the Beat system [5], emphasis points are marked whenever a new incoming word is detected by a language tool (Conexor: www.conexor.fi) and antonyms in the utterance are detected using WordNet [25] to mark contrast points. We can also extend the work by learning the patterns of different head movements or other nonverbal behaviors. Conducting evaluations with human subjects is also necessary to investigate if the head movements generated by the model are perceived to be natural. Finally, we would also like to compare the results of this machine learning approach with the results of our previous rule-based approach or even combine the two approaches to examine if it improves the quality of behaviors generated.

5. ACKNOWLEDGMENTS

We would like to thank Dr. Louis-Philippe Morency for providing technical assistance and helpful comments. 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.

6. REFERENCES

[1] 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.

[2] 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.

[3] J. Carletta. Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation Journal*, 41(2):181–190, 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] N. Cathcart, J. Carletta, and E. Klein. A shallow model of backchannel continuers in spoken dialogue. In *In EACL qŕ03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, pages 51–58, 2003.

[7] E. Charniak. *Statistical Language Learning*. MIT Press, Cambridge, Massachusetts, 1993.

[8] E. Charniak. A maximum-entropy-inspired parser. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, pages 132–139, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[9] D. Decarlo, C. Revilla, M. Stone, and J. J. Venditti. Making discourse visible: Coding and animating conversational facial displays. In *In Proc. Computer Animation 2002*, pages 11–16, 2002.

[10] M. E. Foster and J. Oberlander. Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Language Resources and Evaluation*, 41:305–324(3), 2007.

[11] M. E. Foster, M. White, A. Setzer, and R. Catizone. Multimodal generation in the COMIC dialogue system. In *ACL '05: Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pages 45–48, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[12] U. Hadar, T. J. Steiner, E. C. Grant, and F. C. Rose. Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2:35–46, 1983.

[13] 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.

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

[15] R. W. Hill, J. Belanich, H. C. Lane, M. G. Core, M. Dixon, E. Forbell, J. Kim, and J. Hart. Pedagogically structured game-based training: Development of the elect bilat simulation. In *Proceedings of the 25th Army Science Conference (ASC 2006)*. Association for Computational Linguistics, Noverber, 2006.

[16] B. H. Hwang and L. R. Rabiner. Hidden markov models for speech recognition, August 1991.

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

[18] P. G. Kenny, T. D. Parsons, J. Gratch, A. Leuski, and A. A. Rizzo. Virtual patients for clinical therapist skills training. In C. Pelachaud, J.-C. Martin, E. Andre', G. Chollet, K. Karpouzis, and D. Pele', editors, *IVA*, volume 4722 of *Lecture Notes in Computer Science*, pages 197–210. Springer, 2007.

[19] 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 *IVA*, pages 15–28, 2007.

[20] 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.

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

[22] R. M. Maatman, J. Gratch, and S. Marsella. Natural behavior of a listening agent. In *IVA*, pages 25–36, 2005.

[23] M. Mancini, R. Bresin, and C. Pelachaud. A virtual head driven by music expressivity. *IEEE Transactions on Audio, Speech and Language Processing*, 15(6):1833–1841, 2007.

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

[25] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to WordNet: An on-line lexical database. *Int J Lexicography*, 3(4):235–244, January 1990.

[26] L.-P. Morency, I. de Kok, and J. Gratch. Predicting listener backchannels: A probabilistic multimodal approach. In *IVA*, pages 176–190, 2008.

[27] 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.

[28] R. Nishimura, N. Kitaoka, and S. Nakagawa. A spoken dialog system for chat-like conversations considering response timing. In *TSD*, pages 599–606, 2007.

[29] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[30] M. Strauss and M. Kipp. Eric: a generic rule-based framework for an affective embodied commentary agent. In *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.

[31] W. Swartout, J. Gratch, R. W. Hill, E. Hovy, S. Marsella, J. Rickel, and D. Traum. Toward virtual humans. *AI Mag.*, 27(2):96–108, 2006.

[32] 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.

[33] D. Traum, A. Roque, A. L. P. Georgiou, J. Gerten, B. M. S. Narayanan, S. Robinson, and A. Vaswani. Hassan: A virtual human for tactical questioning. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 71–74, Antwerp, September 2007. Association for Computational Linguistics.

[34] T. Ward and W. Tsukahara. Visual prosody and speech intelligibility in english and japanese. *Pragmatics*, 23:1177–1207, 2004.

[35] A. D. Wilson and A. F. Bobick. Parametric hidden markov models for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):884–900, 1999.