Predicting Speaker Head Nods and the Effects of Affective Information

Jina Lee and Stacy C. Marsella

***Abstract*—During face-to-face conversation, our body is continually in motion, displaying various head, gesture, and posture movements. Based on findings describing the communicative functions served by these nonverbal behaviors, many virtual agent systems have modeled them to make the virtual agent look more effective and believable. One channel of nonverbal behaviors that has received less attention is head movements, despite the important functions served by them. The goal for this work is to build a domain-independent model of speaker's head movements that could be used to generate head movements for virtual agents. In this paper, we present a machine learning approach for learning models of head movements by focusing on when speaker head nods should occur, and conduct evaluation studies that compare the nods generated by this work to our previous approach of using handcrafted rules [1]. To learn patterns of speaker head nods, we use a gesture corpus and rely on the linguistic and affective features of the utterance. We describe the feature selection process and training process for learning hidden Markov models and compare the results of the learned models under varying conditions. The results show that we can predict speaker head nods with high precision (.84) and recall (.89) rates, even without a deep representation of the surface text and that using affective information can help improve the prediction of the head nods (precision: .89, recall: .90). The evaluation study shows that the nods generated by the machine learning approach are perceived to be more natural in terms of nod timing than the nods generated by the rule-based approach.**

***Index Terms*—Embodied conversational agents, emotion, head nods, machine learning, nonverbal behaviors, virtual agents.**

I. Introduction

DURING face-to-face conversation, our body is continually in motion, displaying various facial expressions, head movements, gestures, body postures, and eye gazes. Along with verbal communication, these nonverbal behaviors serve a variety of important functions; they repeat, contradict, substitute, complement, accent, or regulate spoken communication [2]. In addition to these prominent behaviors, the way we use our space (proxemics) or the manners of our haptics are also widely considered as nonverbal communications. Nonverbal behaviors may also be affected by a range of affective phenomena. For example, an angry person might display lowered eyebrows and tensed lips with more expressive body gestures. The behaviors we display not only convey our communicative intent or emotion but also influence the beliefs, emotions, and behaviors of the observers in turn; it is suggested that approximately 60%–65% of the social meanings are derived from our nonverbal cues [3]. The use of appropriate nonverbal behaviors make the interaction more natural and can help create rapport among conversation participants [4].

One of the channels of nonverbal behaviors that has received less attention is head movements. Nevertheless, research has identified a number of important functions served by head movements [5]–[7]. We may nod to show our agreement with what the other is saying, shake to express disapproval and negation, or tilt the head upwards along with gaze aversion when pondering something. As with other nonverbal behaviors, head movements are also influenced by our emotions. Mignault and Chaudhuri [8] 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.

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], [9]–[13], [14]. The incorporation of appropriate head movements in a virtual agent has been shown to have positive effects during human-agent interaction. Virtual agents with natural head motions improve the perception of speech [15] and appropriate head motion not only improves the naturalness of the animation but also enhances the emotional perception of facial animation [16].

Often virtual humans use handcrafted models to generate head movements. For instance, in our previous work, we developed the Nonverbal Behavior Generator (NVBG) [1], which is a rule-based system that analyzes the information from the agent's cognitive processing, such as its internal goals and emotional state as well as the syntactic and semantic structure of the surface text to generate a range of nonverbal behaviors. To craft the rules that specify which nonverbal behaviors should be generated in a 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 [9]–[11], [13] that generate nonverbal behaviors for virtual humans, the NVBG work starts with specific factors that would cause various behaviors to be displayed. Although the knowledge encoded in the NVBG rules has been reused and demonstrated to be effective across

Manuscript received December 01, 2009; revised March 26, 2010; accepted May 10, 2010. Date of current version September 15, 2010. This work was supported in part by the NSF EAPSI fellowship, in part by the Japan Society of for the Promotion of Science (JSPS) fellowship, and in part by the U.S. Army Research, Development, and Engineering Command (RDECOM). The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Nadia Magnenat-Thalmann.

The authors are with the Department of Computer Science, University of Southern California, Los Angeles, CA 90089 USA (e-mail: jlee@ict.usc.edu; marsella@ict.usc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2010.2051874

a range of applications [17]–[20], there are limitations with this approach. One major drawback is that the rules have to be handcrafted. This means that the author of the rules is required to have a broad knowledge of the phenomena he/she wishes to model. However, this is a very challenging task in nonverbal behavior research for several reasons. First, we may not know all the factors that cause people to make certain behaviors. For example, we make head nods to deliver certain communicative functions such as agreement, but we also make head nods that are rhythmic with no clear communicative functions. Secondly, even if we know all the individual factors that can cause the display of nonverbal behaviors, as more and more factors are added, it becomes harder to specify how all those factors contribute to the myriad of behaviors generated. Without complete knowledge of the correlations of the various factors, manual rule construction may suffer from sparse coverage of the rich phenomena.

To complement the limitations of our previous handcrafted literature-based approach (from here we will call this the *rule-based approach*), we present a data-driven, automated approach to derive a model of speaker nonverbal behaviors, which we demonstrate and evaluate. Specifically, the approach uses a machine learning technique—i.e., learning a hidden Markov model (HMM) [21]—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. In particular, we explore the use of features available through shallow parsing and interpretation of the surface text. In addition, we also investigate whether the incorporation of affective information can improve the prediction of the head nod model derived from the surface text. We leave the exploration of deeper features for future work.

There are several advantages with the machine learning approach. First of all, the process is automated. 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. Here, of course, a good understanding of the phenomena is still important as is automated feature selection [22]. Another advantage of this approach is that the learning is flexible and can be customized to learn for a specific context. For example, if we want to learn the head nod patterns for 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 was done in [23]. The advantages of the machine-learning approach makes it a strong alternative to rule-based approach or a substantial enhancement when both are used.

In this paper, we review our prior work on building a domain-independent model of speaker's head movements that can be used to generate head movements for virtual agents [24], [25]. We describe our machine learning approach for learning to predict the speaker's head nods from gesture corpora and also investigate the effect of using the affective sense of utterance during the learning process. Once the patterns of when people nod are learned, we can in turn use the model to generate head nods 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. In addition to this, we also describe the evaluation study conducted with human subjects that compares the head nods predicted by the machine learning model to the nods generated by the rule-based model (NVBG).

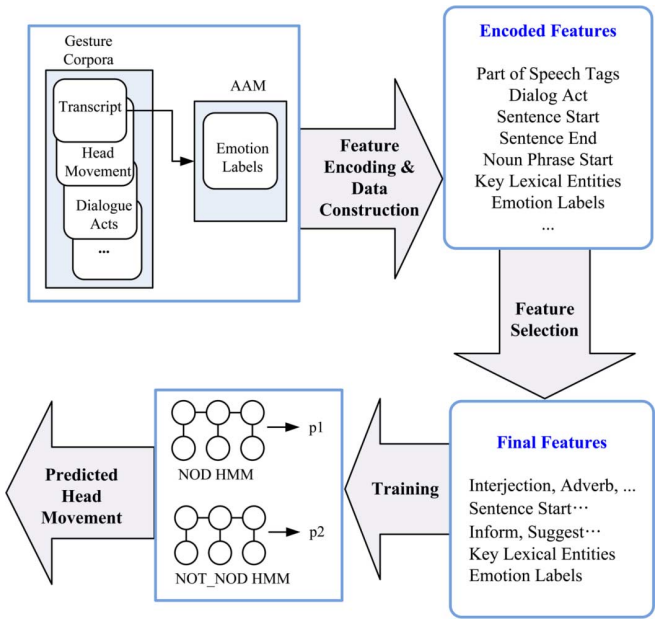


Fig. 1. Overview of the head nod prediction framework. The information in the gesture corpus is encoded and aligned to construct the data set. The feature selection process chooses a subset of the features that are most correlated with head nods. Using these features, probabilistic sequential models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

The following section describes the research on head movements, previous work on modeling head movements for virtual agents, and the approaches each system employs. We then describe our machine learning approach for learning the speaker head nod models in detail. The learning happens in two phases; first we train the model without the affective information, then we retrain the model with affective information. We explain the data construction process, feature selection process, training process, as well as the evaluation of the learned model with test data. Then we explain the incorporation of affective information detected from the utterances and compare the results of the retrained model with the results of the previous model. Fig. 1 shows the overview of the procedures to learn the model (including the incorporation of affective information). The results show that the model is able to predict head nods with high precision and recall rates, but when we incorporate the affective information, the learning can improve even more. Next we describe the evaluation study of the nods generated by the machine learning approach and the rule-based approach. The results show that the nods generated by the machine learning approach are perceived to be more natural in terms of the timing of nod occurrences. Finally, we discuss the results and propose future directions.

II. Related Work

Nonverbal behaviors include everything from facial expressions, gestures, postures, to gaze movements. Researches have identified and categorized a number of functions served

by nonverbal behaviors. Bente and Krämer [26] summarize them into three functional levels: *discourse functions* that are conveyed through pointing or illustrative gestures, *dialogue functions* that regulate the flow of interaction between speaker and listener, and *socioemotional functions* that affect person perception, evaluation, and interaction climate.

Heylen [7] summarizes the functions of head movements during conversations in particular. This includes signalling yes or no, enhancing communicative attention, anticipating an attempt to capture the floor, signalling the intention to continue, marking the contrast with the immediately preceding utterances, and marking uncertain statements and lexical repairs. Kendon [6] 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 [5], McClave describes the linguistic functions of head movements observed from the analysis of videotaped conversations; lateral sweep or head shakes co-occur with concepts of inclusivity such as "everyone" and "everything" and intensification with lexical choices such as "very," "a lot," "great," and "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.

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 [10] verbal/nonverbal behaviors are designed in terms of conversational functions. REA employs head nods to provide feedback and head toss for signalling openness to engage in conversations. BEAT [11] 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.* [9] developed a system for automatic nonverbal generation in which head nod is used as a basic gesture type when listening or is used when no other specific gesture can be suggested. The Nonverbal Behavior Generator [1] generates behaviors given the information about the agent's cognitive processes but also can infer communicative functions from a surface text analysis. Head nods are associated with affirmation, intensification, assumption, and interjections.

Other virtual agents focus on generating expressive behaviors according to the agent's emotional state. Deira [13] 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 [14] is a commentary agent that shows "idle" gestures when no other gestures are requested, but generates various nonverbal behaviors according to its emotional state. Lance and Marsella [27] develop a model of emotionally expressive gaze behaviors based on motion captured data. The model generates gaze behaviors by controlling not only the eye movements but also the head and torso movements according to the emotional state.

Recently, there have been growing efforts to use corpora of nonverbal behavior more extensively. Foster and Oberlander [12] present a corpus-based generation of head and eyebrow motion for virtual agents. To generate behaviors, they either choose the behavior that was most frequently observed or make a weighted choice among all the different behaviors observed for the given context (i.e., feature combination). However, the features they use are based on a specific domain and language tools, which could limit the portability of the model to other systems or domains. For example, the utterances in the corpus are about bathroom tile design and one of the features they use is the user preference of objects being presented (e.g., preference of tile shapes or designers).

Morency *et al.* [22] focus on modeling the *listener's* head movements. They create a model that predicts listener's backchannel head nods using the human speaker's multi-modal features (e.g., prosody, spoken words, and eye gaze). Similarly, other works also use prosodic features to predict listener's backchannel head nod [28]–[30] and [31]. Busso *et al.* [16] use features of actual speech to synthesize emotional head motion patterns for avatars. Their audiovisual database includes recordings of an actress' voice who expresses different emotions and motion caption data that track her head and facial movements. They use prosodic features of the recorded voice and build hidden Markov models to synthesize head motions for each emotional categories and use visual features to generate expressive facial animations. Although the above works build models to generate either the listener's or speaker's head movements, they commonly utilize the speaker's natural speech data, which are shown to have high correlations with head movements [32]. However, in this work where we are focusing on generating head movements for interactive virtual agents, we often do not have the natural speech recordings of the agent's dialog. Many interactive virtual agent systems generate the agent's dialog in real-time according to the human-agent interaction and use automatic text-to-speech (TTS) systems to generate speech, but the generated speech may often sound robotic and unnatural; relying on this information to build a prosody-based model of head movements would introduce many errors.

In the works described above, head nods for virtual agents are either generated based on the findings of the research literature or by learning the patterns from gesture corpora. In the first case, head nods occur to realize communicative functions identified in the literature, but these functions may also be realized through different behaviors other than nods. In addition, it is not necessarily true that the research literature identify all contexts in which head nods occur. In particular, this approach runs a risk of generating less nods than humans and leading the agent to look rigid or unnatural and therefore require addition of ad hoc approaches to embellish the agent's behaviors (see [1] for discussion). Further there are both individual and cultural differences [33], [34] that have only been partially explored in the existing literature. Without modeling these differences, all the virtual agents will nod identically and again make them look unnatural, especially if two or more agents are involved in a multi-party conversation. On the other hand, the machine learning approach can identify more contexts in which nods occur and can capture individual and cultural differences, assuming appropriate data collection and feature selection process. To that end, we hypothesize that the machine learning approach will lead to more natural looking behaviors.

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 available at the time head movements are generated for virtual agents. We also plan to make the model portable to other systems by using features that are easily obtainable across different language tools, instead of relying on information from specific tools [12] or natural speech input [22], [16]. In addition to constructing a framework for predicting head nods, we also investigate whether knowing the affective sense embedded in the utterance can help us. In the following sections, we show that even with shallow model of the surface text, we can learn the model of speaker's head nods with high accuracy and that using the affective sense helps us improve the learning.

III. Predicting Speaker Head Nods

In this section, we describe our machine learning approach for learning the speaker head nods (without affective information). First we describe the gesture corpus used to construct the training data and the feature selection process. We then give a detailed description on how we trained the model and explain the results of the trained model.

A. Gesture Corpus

The AMI Meeting Project is a European-funded multi-disciplinary consortium formed to promote the research of group interaction [35]. The AMI Meeting Corpus is a set of multi-modal meeting records that 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. Fig. 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 is 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 the 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.

Fig. 2. Snapshot of the meeting setting used for AMI meeting corpus [35].

TABLE I
List of Meeting Annotations [35] Used for Learning. Recordings of 17 Meetings Were Used, Which Adds Up to Be Approximately Eight Hours of Annotation

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

- 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.

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

B. Data Alignment and Feature Selection

The main goals for this work are robustness and portability, as well as the ability to generate behaviors using information that is generally available across different systems. One common source of information is the agent's surface text that is generated by the natural language generator. Therefore, for this work, we perform a shallow parsing to analyze the syntactic and semantic structure of the surface string to derive features for the machine learning.

To construct the training data, 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. There are a total of 15 different types of dialogue acts and five different types of head movements. The different dialogue acts are as follows.

Fig. 3. Snapshots of head movements in AMI corpus [35]. From the top: *nod, shake, nodshake*, and *other* head movements.

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

The different head movement types are: nod, shake, nodshake, other, and none. Snapshots of the head movements are shown in Fig. 3. We also obtained the part of speech tags and phrase boundaries (e.g., start/end of verb phrases and noun phrases) by processing the utterances through a natural language parser [36]. In addition, we also combined the features from the Nonverbal Behavior Generator [1]; the nonverbal behavior rules within

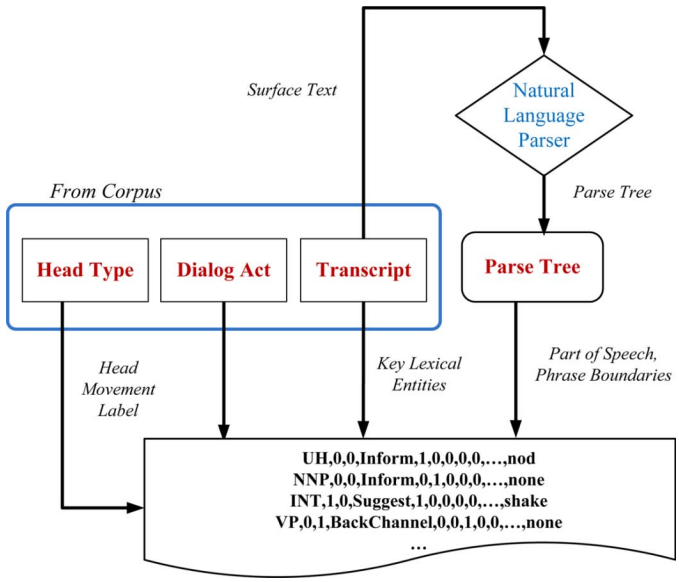


Fig. 4. Data construction process. From the gesture corpus, speaker transcript, dialog act, and head types are extracted. The transcript is sent to the natural language parser to extract the part of speech tags and phrase boundaries. A script automatically cross-references each file to construct the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

TABLE II
KEY LEXICAL ENTITIES. THESE LEXICAL ENTITIES ARE SHOWN TO HAVE HIGH CORRELATIONS WITH HEAD NODS

Affirmation	yeah, yes
Assumption	perhaps, maybe, probably, could, suppose, guess
Intensification	really, very, quite, wonderful, great, absolutely, fantastic, huge, so, amazing

NVBG identify communicative functions and key lexical entities associated with head nods from the psychological research. These communicative functions include *Affirmation*, *Assumption*, and *Intensification*. Table II lists the key lexical entities with the communicative functions served by them and Fig. 4 illustrates the overall data construction process for the prediction model.

From the 17 meeting recordings we used, we collected 10 000 sentences and cross-referenced 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** (e.g., noun, verb, etc.: 29 cases)
- **Dialog act** (each word in the same utterance will have the same dialog act label)
- **Phrase boundaries** (sentence start/end, noun phrase start, verb phrase start)
- **Key lexical entities** (keywords known to be highly associated with nods from psychological research)

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, a subset of the available features was selected by the frequency of head nods that occurred with each feature. Table III lists the frequency counts of

TABLE III
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	
Interjection	427 (16%)
Proper Noun	300 (12%)
Conjunction	239 (9%)
Adverb	238 (9%)

Dialog Act	
Inform	910 (35%)
BackChannel	387 (15%)
Suggest	265 (10%)

Phrase Boundaries	
sentence_start	2268 (88%)
np_start	493 (19%)
vp_start	391 (15%)

NVBG Rule	
key_lexical_entities	594 (23%)

TABLE IV
FINAL FEATURES SELECTED FOR TRAINING. THE FEATURES WERE SELECTED BASED ON THE RESULTS OF TABLE III. THE LABEL "REMAINDER" INCLUDES EVERYTHING NOT FALLING UNDER OTHER CATEGORIES. THE EMOTION LABEL WAS USED FOR THE SECOND PHASE OF LEARNING DESCRIBED IN SECTION IV

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
Emotion Label	Anger, Disgust, Fear, Guilt, Interest, Joy, Sadness, Shame, Surprise, Neutral
Key Lexical Entities	y, n

these features (out of 2590 words with nods). It shows that head nods occurred more frequently at the beginnings of utterances and noun/verb phrases than at the end. 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*. Based on the results described above, the final features were selected for training. Table IV lists the final features used for training the models.

C. Training Process

To learn the head nod model, HMMs [21] were trained. HMMs are statistical models that are widely used for learning patterns where a sequence of observations is given. Some of the applications where HMMs have been successfully used are gesture recognition, speech recognition, and part-of-speech tagging [37]–[39]. 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 sequences of three words to form a set of trigrams. We constructed trigrams because unigrams or bigrams are not long enough to learn the dynamics or the "context" of the observation sequence, and observation sequences longer than four words would result in sparse data since a large number of head nods in the gesture corpus span over fewer than four words. The constructed trigrams that formed our data set are overlapping, so that each word is part of three exactly trigrams. 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 labeled as a nod instance, and the same applied for other head movement types.

TABLE V
MEASUREMENTS FOR THE PERFORMANCE OF THE LEARNED MODELS

Measurement	**Equation**	**Value**
Accuracy	$\frac{tp+fn}{tp+fp+tn+fn}$	.8577
Precision	$\frac{tp}{tp+fp}$	.8366
Recall	$\frac{tp}{tp+fn}$	.8890
F-score	$\frac{2\times precision\times recall}{precision+recall}$	.8620

TABLE VI
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 V

	Precision	Recall	F-score
Adverb	-0.0108	+0.0128	+0.0010
Verb Phrase Start	+0.0496	+0.0190	+0.0350
Noun Phrase Start	+0.0578	+0.0128	+0.0343
Suggest	+0.0543	-0.0056	+0.0269
Interjection	+0.0640	+0.0006	+0.0331

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 movements types other than a nod or no head movement. 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. There were a total of 1318 instances of "nod" trigrams.

After collecting all the "nod" trigrams, 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 settings of HMM (i.e., the number of hidden states) that produce the best result, we performed a ten-fold cross-validation for each parameter setting. That is, we split the remaining 80% of the data into ten parts and used one part as a validation set and nine parts as a training set. After training the model, we obtained the performance measurements of the model. We repeated this process ten times and obtained an average performance measurement for each given number of hidden states. We then determined the best number of hidden states by comparing the average performance measurements. After this, we combined all ten parts and trained the final "NOD HMM" with the chosen number of hidden states. In our case, the model with six hidden states performed the best. Similarly, we collected the positive instances of "NOT_NOD" trigrams (i.e., trigrams with head movements other than nod or no head movements, total 72 344 trigrams) and repeated the above steps to

TABLE VII
MEASUREMENTS FOR THE PERFORMANCE OF THE LEARNED MODEL. FOR THE ONE-SIDED T-TESTS, * MARKS $P < .05$ AND ** MARKS $P < .01$

	Model	Accuracy	Precision	Recall	F-score
RESULT	No Emo	.8577	.8366	.8890	.8620
	Emo Word	.8577	.8348	.8926	.8626
	Emo Sentence	.8963	.8939	.8994	.8966
T-TEST (one-tailed)	No Emo vs. Emo Word	p=.498	p=.395	p=.154	p=.427
	No Emo vs. Emo Sentence	p<.001**	p<.001**	p=.001**	p<.001**
	Emo Word vs. Emo Sentence	p<.001**	p<.001**	p=.023*	p<.001**

train a final "NOT_NOD HMM." Finally, we ran the test set (20% of the entire data left out which were not used in training) through the "NOD HMM" and "NOT_NOD HMM" and classified each sample to have the head movement of whichever model produced a higher probability. The testing was also done in ten-folds to reduce variability.

D. Results and Discussion

To measure the performance of our learned model, we computed the accuracy, precision, recall, and F-score of the learned model. Accuracy is the ratio of samples that were correctly classified. Precision is the ratio of the number of nods predicted by the learned model which are actually nods to the total number of nods predicted by the data. Recall is the ratio of the number of nods predicted by the learned model which are actually nods to the number of nods in the actual data. For the F-score, we gave the recall and precision the same weight (F_1). Table V summarizes the results. We achieved .8577 for accuracy, .8366 for precision, .8890 for recall, and .8620 for F-score rates.

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). An insight into where the model failed could be explained by looking at false positive and false negative samples. A majority of false positives (predicted nods that are not actually nods) occurred on trigrams that represented the end of sentences. On the other hand, false negatives (actual nods that were missed by the model) occurred largely when the trigram included words that are the beginning of verb phrases whose parts of speech were marked as "Remainder."

In addition to the main results presented in Table V, a second 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 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 VI. Specifically, *Adverb* affected the learning very marginally when taken out, whereas removing *Verb Phrase Start*, *Noun Phrase Start*, *Suggest*, and *Interjection* resulted in a larger change in both precision and F-score values. Interestingly, in the case of *Verb Phrase Start* and *Noun Phrase Start*, our previous rule-based approach (NVBG) inserted head nods in those places to make the agent look more lifelike, because the NVBG was under-generating behaviors. The results of the second experiment raise a need for a more sophisticated automatic feature selection process such as the method used by Morency *et al.* [22], 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.

IV. USING AFFECTIVE INFORMATION

The previous section described the framework for the speaker head nod prediction model using the information obtained from shallow parsing of the surface text. In this section, we incorporate the affective information and investigate whether it can improve the learning. We use the Affect Analysis Model (AMM) [40] to detect the affective sense of each word as well as each sentence and retrain the model. The results of the model using affective information over each word and over each sentence are compared to the previous model using no affective information.

A. Affect Analysis Model

The AAM [40] is a rule-based system aimed for the recognition of ten emotions from text. Given a sentence, AAM performs a series of analysis and produces an emotion vector that represents the intensities of ten emotion categories: Anger, Disgust, Fear, Guilt, Interest, Joy, Sadness, Shame, Surprise, and Neutral. The initial stages of the analysis test for occurrences of emoticons, abbreviations, acronyms, or punctuation marks and parses the sentence through a syntactical parser to obtain more exhaustive information of the given sentence. AAM then goes through a word-level analysis in which the words in the sentences are looked up against a database that contains a set of emotional words with their intensity values. If the sentence includes modifiers (e.g., "very," "extremely," "hardly," etc.), the emotional intensities of the modified words are adjusted accordingly. After the word-level analysis, AAM analyzes the types of phrases contained in the sentence (e.g., adjective/noun/verb plus noun phrase) and the syntactical structure of the sentence (e.g., simple/complex/compound sentences) to further modify the intensities of each emotion category. The evaluation of AAM shows that the system's output of the dominant emotion label of sentences agreed 79.4% of the time with at least one out of three human annotators and 70% of the time with at least two annotators [40].

B. Retraining of the Model and Results

For our second phase of learning, we obtained the emotion vectors of each word in the sentence as well as that of the whole sentence from the Affect Analysis Model. The most dominant

emotion category in the emotion vector was used as the emotion label of each word or sentence. To learn the head nod models, we retrained the model first using emotion label of each word then the emotion label of each sentence as additional features. The training process is identical to that described in Section III-C. The number of states for the final trained model with no emotion, emotion over word, and emotion over sentence were 6, 2, and 3, respectively.

Table VII summarizes the results, including one-tailed t-tests among the three conditions. When emotion label for each word was used, there were marginal changes in the precision, recall, and F-score rates compared to when no affective information was used, and these changes were not statistically significant. Therefore, using the affective information over words as an additional feature had very little impact on learning. On the other hand, when the affective information for the whole sentence was used, the accuracy, precision, recall, and F-score rates all increased compared to the other two models, with the precision rate showing the greatest increase by .0573 (when compared to the model with no affective information). The t-tests show that all of these increases were statistically significant.

A closer look at the training output tells us that when the affective information over sentences was used, there were 46% fewer false positives (i.e., number of predicted nods that are not actually nods) compared to when no affective information was used, resulting in a large increase in the precision rate. More specifically, the model predicted fewer false nods at the end of the sentences. Arguably, this increase in the precision rate by lowering the number of false positives is especially important because predicting nods at times when they should not occur could lead to false implicature. For example, the model could be generating a nod that emphasizes a wrong point in the utterance.

There are also several possible explanations for why using the affective information over sentences outperforms using the affective information over words. First, it may be that the Affect Analysis Model performs better on sentences than on words. To produce an emotion label over a word, AAM simply looks up the word in the database, whereas for sentences, it goes through a more sophisticated analysis. Secondly, nods may need a wider context. Specifically, they can have an association with higher level semantic or pragmatic factors, which can span over phrases or sentences than a single word. This emphasizes a deeper analysis to improve the learning.

V. Evaluation Study

In this section, we describe an evaluation study with human subjects to compare the perception of head nods driven by our rule-based approach, machine learning approach, and real human behavior. To do this, an online evaluation study was conducted. Here we especially stress the importance of conducting evaluation studies with human subjects because despite the fact that our machine learning model predicts head nods with high precision, recall, and f-score values, it does not directly guarantee that the predicted nods will also look natural to human eyes.

A. Hypothesis

Our main interest for the evaluation study was to investigate the perception of head nods generated by the two approaches as well as the nods made by real humans but displayed through a virtual agent. To answer this, we compared the following schemes for generating head nods:

- head nods made by humans and displayed through a virtual agent;
- head nods generated by the machine learning approach and displayed through a virtual agent;
- head nods generated by the rule-based approach and displayed through a virtual agent.

In this study, we hypothesized that

> The occurrence of nods generated by the machine learning approach will be perceived to be more natural than the nods generated by the rule-based approach.

We based the hypothesis from the fact that because machine learning 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, the handcrafted rule-based approach largely ignores the context of the interaction when generating behaviors, so it may over-generate or under-generate behaviors. In addition to the main hypothesis, we also expected that the human-made nods would look more natural than nods from either the machine learning approach or rule-based approach.

B. Generating Nods From the Machine Learning Model

To generate head nods for a given sentence using the machine learning model, we used the trained model that incorporates affective information over the whole sentence. Since the machine learning model predicts head nod occurrences on a trigram-level, we need a way to determine where the nod occurs in the trigram. To do this, for each word in the sentence, we looked at all the trigrams that particular word is part of (there are three in total). Similar to the majority voting method we used for determining the head movement label for the trigrams in the training process, if two out of three trigrams the particular word is involved were classified as nods, that word was predicted to accompany a nod.

C. Evaluation Study Methods

1) Participants: Twenty-nine participants were recruited via e-mail and web postings. There were 25 males and four females with ages ranging from 19 to 41 ($\mathrm{M} = 28.9$ years, $\mathrm{SD} = 4.04$ years).

2) Stimuli: We created video clips of a virtual agent displaying head nods while speaking an utterance. Fig. 5 shows a snapshot of the video clip. We randomly selected seven utterances spoken by several individuals from the gesture corpus used in the machine learning 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 machine learning model to obtain head nod predictions.

With the nod predictions from both the rule-based approach and the machine learning approach, we created three versions of video clips for each utterance: head nods displayed by the human in the gesture corpus, head nods generated by the machine learning approach, and head nods generated by the rule-based approach. Therefore, there were a total of 21 video clips (7 utterances $\times$ 3 conditions). In all three conditions, the magnitude, velocity, and length of the nods were unified; the models

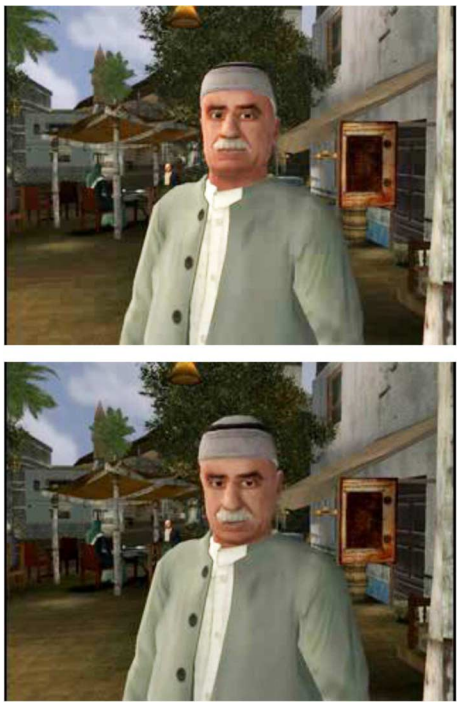


Fig. 5. Snapshot of the video clip shown in the evaluation study.

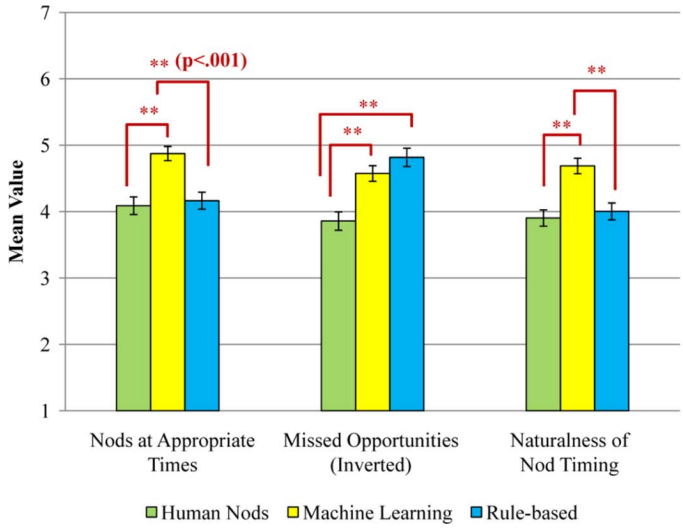


Fig. 6. Mean values for the evaluation study. Vertical bars denote the 95% confidence intervals and asterisks (**) mark statistical significance with $p < .001$.

only predicted the timing of the nods and not the dynamics. Therefore, the only differences among the different conditions were the frequency of the nods and the timing of when nods occurred. 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, machine learning approach, and rule-based approach were 3.14, 5.29, and 4.5.

3) 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. Following this, participants went through seven sets (one for each utterance) of video evaluations, each consisting of three video clips for each condition. The order of the evaluation sets and the video clips representing each condition were all randomized. The video clips lasted about 10 s. After watching each video, users were asked to answer questions on various aspects of the head nod timings. The specific questions asked were

1) Do the agent's nods occur at appropriate times?
2) Were there times when the agent should have nodded but did not?
3) How natural is the timing of nod occurrences overall?

Participants answered the questions using a scale from 1 to 7 (1 meaning "Never" or "Not natural at all" and 7 meaning "Always" or "Very Natural").

D. Results

The analysis 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. For the second question, we inverted the values so that the higher values indicate better results. Fig. 6 shows the mean values for the three questions. In all three questions, the results show statistical significance.

For the perception of nods at appropriate times (Question 1), there was a significant effect of the generation approach $[F(1.82,\ 293.15) = 14.882, p < .001]$. In general, participants perceived that nods generated by the machine learning approach $(M = 4.877, SE = 0.107)$ had more cases of nods at appropriate times, followed by nods generated by rule-based approach $(M = 4.167, SE = 0.128)$ and human nods $(M = 4.093, SE = 0.132)$. The pairwise comparisons show that there is a significant difference between human nods and machine learning approach, and between machine learning approach and rule-based approach.

For the perception of missed head nod opportunities (Question 2), there was also a significant effect of the generation approach $[F(1.84,\ 296.23) = 15.908, p < .001]$. Participants perceived that the rule-based approach $(M = 4.815, SE = 0.138)$ missed less nod opportunities followed by machine learning approach $(M = 4.458, SE = 0.117)$ and human nods $(M = 3.858, SE = 0.138)$. The pairwise comparisons show that there is a significant difference between human nods and machine learning approach, and between human nods and rule-based approach.

Finally, there was also a significant effect of the generation approach on the overall naturalness of the nod timing (Question 3) $(F(1.80,\ 289.47) = 13.632, p < .001)$. Participants rated the nods generated by machine learning $(M = 4.691, SE = 0.116)$ approach the highest followed by nods generated by rule-based approach $(M = 4.006, SE = 0.127)$ and human-made nods $(M = 3.907, SE = 0.123)$. Similar to the results of Question 1, there is a significant difference between human nods and machine learning approach, and between machine learning approach and rule-based approach.

E. Discussion

The results of the evaluation study show that there was a significant effect of the generation approach on the ratings for nods at appropriate times, missed nod opportunities, and naturalness of the nod timing. Furthermore, there were significant differences between the machine learning approach and the rule-based approach on the perception of nods at appropriate times

and the overall naturalness of nod timings. Therefore, we can conclude that our hypothesis was partly validated.

Additionally, the machine learning approach produced better results than human nods in all three questions with statistical differences, which is in opposition to our expectation. A possible explanation for this result could come from the fact that the machine learning model is a general model trained on the data from many different people. Therefore, the model predicts where people will most likely nod on average. On the other hand, each video showing human nods was based solely on nodding behaviors of the differing individuals who originally spoke the different utterances. Thus, the participants are comparing a particular person's nods on a particular utterance to those from a general model learned from many people's nods on many different utterances. Since each individual style 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. This result sheds light on an important fact; there is a wide variability in human nonverbal behavior. Some people nod quite often, for example, while others rarely do. Whereas an individual's behavior is a baseline in the sense that we seek to achieve at least that level of performance in our models, an individual's behavior may not necessarily be the "gold standard" of what will be perceived as the most natural behavior.

VI. Conclusions and Future Direction

In this paper, we presented an approach to learning a probabilistic model to predict head nods using a gesture corpus. In this approach, 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. Using these features, 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. We also explored the effect of affective information to improve the learning. We detected the affective sense over words and sentences and incorporated them during the learning process. Comparing the results of three different models (no emotion, emotion over word, emotion over sentence), it is shown that we can improve the prediction of speaker's nods when using the affective information of the whole sentence. Finally, we conducted an evaluation study with human subjects on the perception of nods generated by our previous rule-based approach, machine learning approach, and real human nods. The results show that timings of nods generated by the machine learning approach were perceived to be more natural than the nods generated by the rule-based approach.

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 the 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. Knowing the affective sense of the utterance is also shown to be helpful in learning the models, but a simple lookup to determine the emotion for the words can damage the learning, which emphasizes the need for a deeper analysis to improve the learning. The evaluation study shows that the machine learning approach captures the "naturalness" of the nodding behaviors, in terms of nod timing, better than the rule-based approach, which often ignores the context of the interaction.

This work could be extended in several ways. First of all, as mentioned earlier, we can use a more sophisticated feature selection process and retrain the models. The current feature selection process simply looks at the co-occurrence of each feature and head nods, but we can also investigate the correlations among different features and their impacts on head nods. Secondly, we can use similar approach to learn patterns of different head movements and other nonverbal behaviors, or to learn patterns of behaviors across individuals and cultures. Thirdly, we may extend the work to include deeper features in addition to the linguistic features. For example, we may look at the concurrence of other behaviors to see if there is a correspondence with head nods or encode the social context, such as the dominance relationship between the conversation participants, as additional features. We also plan to extend the evaluation study with human subjects. We are especially interested in evaluating what the subjects infer from looking at the generated behaviors, including whether the agent looks trustworthy and friendly, or whether the agent's emphasis points in the utterances coincide with those perceived by the users.

References

[1] J. Lee and S. Marsella, "Nonverbal behavior generator for embodied conversational agents," in *Proc. 6th Int. Conf. Intelligent Virtual Agents*, 2006, pp. 243–255, Springer.

[2] M. Knapp and J. Hall, *Nonverbal Communication in Human Interaction*, 4th ed. Fort Worth, TX: Harcourt Brace College, 1997.

[3] J. K. Burgoon, , M. L. Knapp and G. R. Miller, Eds., "Nonverbal signals," in *Handbook of Interpersonal Communication*, 2nd ed. Thousand Oaks, CA: Sage, 1994, pp. 229–285.

[4] L. Tickle-Degnen and R. Rosenthal, "The nature of rapport and its nonverbal correlates," *Psychol. Inq.*, vol. 1, no. 4, pp. 285–293, 1990.

[5] E. Z. McClave, "Linguistic functions of head movements in the context of speech," *J. Pragmat.*, vol. 32, pp. 855–878, 2000.

[6] A. Kendon, "Some uses of the head shake," *Gesture*, vol. 2, pp. 147–182, 2002.

[7] D. Heylen, "Challenges ahead: Head movements and other social acts in conversations," in *Proc. Social Presence Cues Symp (AISB 2005)*, 2005.

[8] A. Mignault and A. Chaudhuri, "The many faces of a neutral face: Head tilt and perception of dominance and emotion," *J. Nonverb. Behav.*, vol. 2, no. 27, pp. 111–132, Jun. 2003.

[9] W. Breitfuss, H. Prendinger, and M. Ishizuka, "Automated generation of non-verbal behavior for virtual embodied characters," in *Proc. 9th Int. Conf. Multimodal Interfaces (ICMI'07)*, New York, 2007, pp. 319–322, ACM.

[10] J. Cassell, "More than just another pretty face: Embodied conversational interface agents," *Commun. ACM*, vol. 43, pp. 70–78, 2000.

[11] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore, "BEAT: The behavior expression animation toolkit," in *Proc. 28th Annu. Conf. Computer Graphics and Interactive Techniques (SIGGRAPH'01)*, New York, 2001, pp. 477–486, ACM.

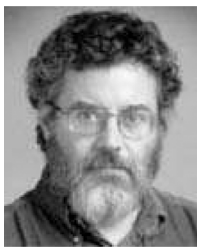
[12] M. E. Foster and J. Oberlander, "Corpus-based generation of head and eyebrow motion for an embodied conversational agent," *Lang. Resour. Eval.*, vol. 41, pp. 305–324, 2007.

[13] F. L. A. Knoppel, A. S. Tigelaar, D. O. Bos, T. Alofs, and Z. Ruttkay, "Trackside DEIRA: A dynamic engaging intelligent reporter agent," in *Proc. 7th Int. Joint Conf. Autonomous Agents and Multiagent Systems (AAMAS'08)*, Richland, SC, 2008, pp. 112–119, International Foundation for Autonomous Agents and Multiagent Systems.

[14] M. Strauss and M. Kipp, L. Padgham, D. C. Parkes, J. Mller, and S. Parsons, Eds., "Eric: A generic rule-based framework for an affective embodied commentary agent," in *Proc. 7th Int. Joint Conf. Autonomous Agents and Multiagent Systems (AAMAS'08)*, Richland, SC, 2008, pp. 97–104, International Foundation for Autonomous Agents and Multiagent Systems.
[15] 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," *Psychol. Sci.*, vol. 15, pp. 133–137, Feb. 2004.
[16] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, "Rigid head motion in expressive speech animation: Analysis and synthesis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1075–1086, Mar. 2007.
[17] W. Swartout, J. Gratch, R. W. Hill, E. Hovy, S. Marsella, J. Rickel, and D. Traum, "Toward virtual humans," *AI Mag.*, vol. 27, no. 2, pp. 96–108, 2006.
[18] 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 *Proc. 8th SIGdial Workshop Discourse and Dialogue*, Antwerp, Belgium, Sep. 2007, pp. 71–74, Association for Computational Linguistics.
[19] P. G. Kenny, T. D. Parsons, J. Gratch, A. Leuski, and A. A. Rizzo, , C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, Eds., "Virtual patients for clinical therapist skills training," in *IVA*, ser. Lecture Notes in Computer Science. New York: Springer, 2007, vol. 4722, pp. 197–210.
[20] 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 *Proc. 25th Army Science Conf. (ASC 2006)*, Nov. 2006, Association for Computational Linguistics.
[21] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE* vol. 77, no. 2, pp. 257–286, Feb. 1989.
[22] L.-P. Morency, I. de Kok, and J. Gratch, "Predicting listener backchannels: A probabilistic multimodal approach," in *Proc. 8th Int. Conf. Intelligent Virtual Agents*, 2008, pp. 176–190.
[23] 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 *Proc. 7th Int. Conf. Intelligent Virtual Agents*, 2007, pp. 15–28.
[24] J. Lee and S. Marsella, "Learning a model of speaker head nods using gesture corpora," in *Proc. 8th Int. Joint Conf. Autonomous Agents and Multiagent Systems (AAMAS'09)*, 2009, International Foundation for Autonomous Agents and Multiagent Systems.
[25] J. Lee, A. Neviarouskaya, H. Prendinger, and S. Marsella, "Learning models of speaker head nods with affective information," in *Proc. 3rd Int. Conf. Affective Computing and Intelligent Interaction (ACII'09)*, 2009, International Foundation for Autonomous Agents and Multiagent Systems.
[26] G. Bente and N. Krämer, "Virtual gestures. Embodiment and nonverbal behavior in computer-mediated communication," in *Emotion in the Internet*, A. Kappas, Ed. Cambridge, U.K.: Cambridge Univ. Press, to be published.
[27] B. Lance and S. Marsella, "Glances, glares, and glowering: How should a virtual human express emotion through gaze?," *J. Autonom. Agents and Multi-Agent Syst.*, vol. 20, no. 1, pp. 50–69, 2010.
[28] T. Ward and W. Tsukahara, "Visual prosody and speech intelligibility in English and Japanese," *Pragmatics*, vol. 23, pp. 1177–1207, 2004.
[29] R. Nishimura, N. Kitaoka, and S. Nakagawa, "A spoken dialog system for chat-like conversations considering response timing," TSD 2007, pp. 599–606.
[30] N. Cathcart, J. Carletta, and E. Klein, "A shallow model of backchannel continuers in spoken dialogue," in *Proc. 10th Conf. Eur. Chapter of the Association for Computational Linguistics (EACL '03)*, 2003, pp. 51–58.
[31] R. M. Maatman, J. Gratch, and S. Marsella, "Natural behavior of a listening agent," in *Proc. 5th Int. Conf. Intelligent Virtual Agents*, 2005, pp. 25–36.
[32] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, "Visual prosody: Facial movements accompanying speech," in *Proc. AFGR*, 2002, pp. 381–386.
[33] D. Matsumoto, "American-Japanese cultural differences in the recognition of universal facial expressions," *J. Cross-Cultural Psychol.*, vol. 23, pp. 72–84, 2002.
[34] S. K. Maynard, "Interactional functions of a nonverbal sign: Head movement in Japanese dyadic casual conversation," *J. Pragmat.*, vol. 11, pp. 589–606, 1987.
[35] J. Carletta, "Unleashing the killer corpus: Experiences in creating the multi-everything AMI meeting corpus," *Lang. Resour. Eval. J.*, vol. 41, no. 2, pp. 181–190, 2007.
[36] E. Charniak, "A maximum-entropy-inspired parser," in *Proc. 1st Conf. North Amer. Chapter of the Association for Computational Linguistics*, San Francisco, CA: Morgan Kaufmann, 2000, pp. 132–139.
[37] A. D. Wilson and A. F. Bobick, "Parametric hidden Markov models for gesture recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 9, pp. 884–900, Sep. 1999.
[38] B. H. Hwang and L. R. Rabiner, "Hidden Markov models for speech recognition," *Technometrics*, vol. 33, no. 3, pp. 251–272, 1991.
[39] E. Charniak, *Statistical Language Learning*. Cambridge, MA: MIT Press, 1993.
[40] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Textual affect sensing for sociable and expressive online communication," in *Proc. 2nd Int. Conf. Affective Computing and Intelligent Interaction (ACII'07)*, Berlin, Germany: Springer-Verlag, 2007, pp. 218–229.

Jina Lee received the B.S. degree in computer science from the University of Washington, Seattle, in 2004 and the M.S. degree in computer science from University of Southern California (USC), Los Angeles, in 2006. She is currently pursuing the Ph.D. degree in computer science at the Institute for Creative Technologies at USC.

Her research interests are in human-computer interaction, statistical learning, and cognitive science. Her current research includes developing models of nonverbal communication and interaction, which informs the synthesis of nonverbal behaviors for virtual agents through multimodal interfaces.



Stacy C. Marsella received the Ph.D. degree in computer science from Rutgers University, New Brunswick, NJ, in 1993.

He is the Associate Director of Social Simulation Research at the University of Southern California's (USC), Los Angeles, Institute for Creative Technologies (ICT), a Research Associate Professor in the Department of Computer Science, and co-director of USC's Computational Emotion Group. He works on computational models of human behavior, including emotion, cognition, and social behavior. He also works on incorporating these models into virtual humans, autonomous characters that look human, act like humans, and can interact with humans within virtual worlds. He has extensive experience in the design and construction of simulations of social interaction for a variety of research and educational applications. He has published over 150 technical articles.

Dr. Marsella is an Associate Editor of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, on the editorial board of the *Journal of Experimental & Theoretical Artificial Intelligence*, and on the steering committee of the Intelligent Virtual Agents conference. He is a member of the Association for the Advancement of Artificial Intelligence (AAAI), the International Society for Research on Emotions (ISRE), and the Society of Experimental Social Psychology.