Learning Models of Speaker Head Nods with Affective Information

Jina Lee
University of Southern California, USA
jlee@ict.usc.edu

Alena Neviarouskaya
University of Tokyo, Japan
lena@mi.ci.i.u-tokyo.ac.jp

Helmut Prendinger
National Institute of Informatics, Japan
helmut@nii.ac.jp

Stacy Marsella
University of Southern California, 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, and may also be influenced by our emotions. The goal for this work is to build a domain-independent model of speaker's head movements and investigate the effect of using affective information during the learning process. Once the model is learned, it can later be used to generate head movements for virtual agents. In this paper, we describe our machine-learning approach to predict speaker's head nods using an annotated corpora of face-to-face human interaction and emotion labels generated by an affect recognition model. We describe the feature selection process, training process, and the comparison of results of the learned models under varying conditions. The results show that using affective information can help predict head nods better than when no affective information is used.

1. Introduction

During face-to-face conversation, the head is constantly in motion, especially during speaking turns [7]. These movements are not random; research has identified a number of important functions served by head movements [16] [11] [8] [9]. 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 to express disapproval and negation, or tilt the head upwards along with gaze aversion when pondering something. Head movements are also influenced by our emotions. For example, Mignault and Chaudhuri [17] 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 roles that head movements play in human-human interaction, virtual agent systems have incorporated head movements to realize a variety of functions [1] [4] [12] [13] [18]. The incorporation of appropriate head movements in a virtual agent has been shown to have positive effects during human-agent interaction [19]. 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. In addition, we would like to investigate whether the use of affective information helps us during the learning process. To make the model compatible 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.

Prior work [14] presented and evaluated an approach for learning to predict speaker head nods from gesture corpora using machine learning techniques. Once learned, such a model could be used to generate head movements for virtual agents. In particular, the learned model could be incorporated into a larger system like Nonverbal Behavior Generator [13]. In addition, we can use the same machine-learning techniques to learn the head nod patterns for specific context such as different cultures, personalities, or individuals.

In this paper we ask the question whether the incorporation of affective information can improve the prediction of the head nod model. To address this question, we extend our prior work to incorporate affective information inferred from the utterance using the Affect Analysis Model [20].

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 under different conditions. The results show that using affective information can help predict head nods better than when no affective information is used dur-

978-1-4244-4799-2/09/$25.00

ing learning. 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 [8] [9] [11] [16]. Heylen [9] 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 [11] 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 [16], 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,' and 'really.' Side-to-side shakes also correlate with expressions of uncertainty and lexical repairs. In general, head nods are used as an affirmative or positive signal, showing one's understanding, approval, and agreement, whereas shakes are used to express valenced intents that convey affective information, such as disapproval or disbelief. However, nods could also be used during negative emotions, such as more pronounced vertical head movements during anger.

Following the studies on nonverbal behaviors, many virtual agents model these behaviors. Some generate the behaviors according to the 'conversation phenomena' or discourse structure. BEAT [4] generates eyebrow flashes and beat gestures when the agent describes a new object. In the system developed by Breitfuss et al. [1], head nod is used as a basic gesture type for listener and when no other gesture is suggested. In our previous work [13], we developed the Nonverbal Behavior Generator that implements a set of rules that map abstract communicative functions such as expression of intensification or affirmation to specific gestures including head nods.

Other systems focus on generating expressive behaviors according to the agent's emotional state. Mancini et al. [15] use acoustic cues and emotions to show how musical expressivity could be transformed to behavioral expressivity through head movements. Deira [12] and ERIC [22] are reporter agents that generate more pronounced movements as the agent's excitement level rises during the report. Busso et al. [2] 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 through an animated face. Their evaluation shows that head motion modifies the emotional perception of facial animation, especially in valence and activation domain.

As noted earlier, we modeled the speaker head movements using human gesture corpora, relying on the linguistic features of the surface text [14]. The results show that human head nods could be predicted with high performance measures using machine learning approach even without a rich markup of surface text. In this paper, we want to investigate the use of affective information during the learning of speaker head nod models. We perform this by using the detected emotion label of each word in the surface text as well as the emotion label over the whole sentence during training process. Finally, we plan to make the model portable to other systems by using features such as part of speech tags that are easily obtainable across different language tools. In the following section, we show that even with a shallow model of the surface text and the use of emotion label over the whole sentence, 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. Figure 1 shows the overview of the procedures to learn the model. First we describe the gesture corpus used to constructed the training data and the Affect Analysis Model [20], from which we obtained the emotion labels for each utterance. Then we explain 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

The AMI Meeting Corpus [3] is a set of multi-modal 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. The meeting records could be either from scenario meetings, in which participants play the roles of employees in an electronics company, or from non-scenario meetings where participants are colleagues from the same area and have discussions on their research topics. In either case, 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

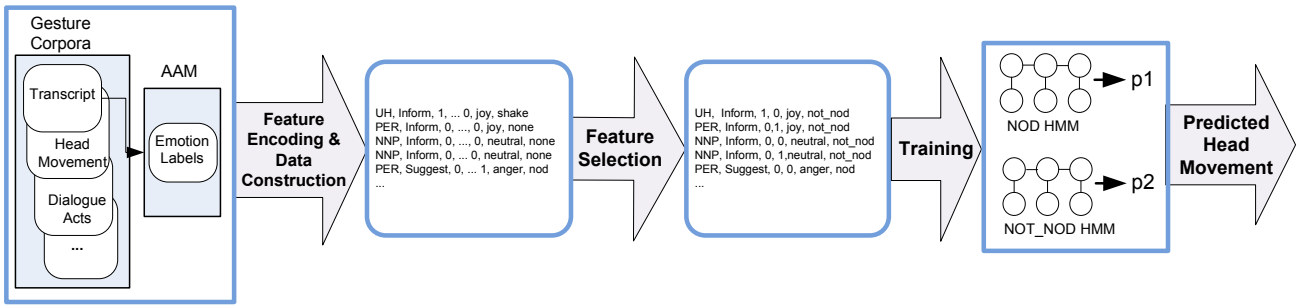

Figure 1. Overview of the head nod prediction framework. The information in the gesture corpus is encoded and aligned to construct the data set. The feature selection process chooses a subset of the features that are most correlated with head nods. Using these features, probabilistic sequential models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

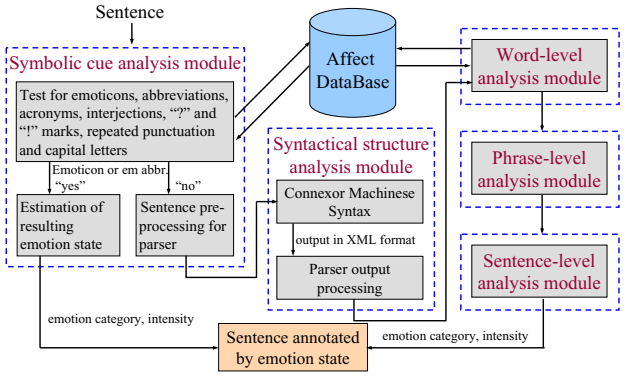

Figure 2. Working Flow of Affect Analysis Model

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.

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 annotations.

3.2. Affect Analysis Model

The Affect Analysis Model (AAM) [20] is a rule-based system aimed for the recognition of ten emotions from text. The emotion labels are: anger, disgust, fear, guilt, interest, joy, sadness, shame, surprise, and neutral. The algorithm for the analysis of affect consists of five stages, as shown in Figure 2.

During the *Symbolic Cue Analysis*, each sentence is tested for occurrences of emoticons, abbreviations, acronyms, interjections, '?' and '!' marks, repeated punctuation and capital letters. We defined a set of rules for cases when single or multiple symbolic cues occur in the sentence. Next in the *Syntactical Structure Analysis* level, non-emotional abbreviations and acronyms are replaced by their proper transcriptions before sent to the parser. AAM uses a syntactical parser, Connexor Machinese Syntax (http://www.connexor.com), which returns exhaustive information for each sentence. From the parser output we can read off the characteristics of each token and the relations between them in a sentence (e.g. subject, verb, object, and their attributes).

In the *Word-level Analysis* stage, the database is examined for presence of analyzed words. The affective features of an emotional word found in the database are represented as a vector of emotional state intensities e = [anger, disgust, sadness, fear, guilt, interest, joy, shame, surprise] (e.g. e = [0.2, 0, 0.7, 0, 0, 0, 0, 0, 0] for the word 'frustrated'). In case of modifiers, the system identifies its coefficient and for adjectives in comparative or superlative forms, the intensities of the emotional vectors are multiplied by 1.2 or 1.4, respectively.

During the *Phrase-level Analysis*, AAM detects emotions involved in phrases. We defined general types of phrases and rules for processing them with regard to affective content as the following: (1) adjective phrase: modify the vector of adjective; (2) noun phrase: output vector with the maximum intensity within each corresponding emotional state in analysing vectors; (3) verb plus noun phrase: if verb and noun phrase have opposite valences, we consider the vector of the verb as dominant; if valences are the same, output vector with maximum intensities in cor-

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 1. Types of dialog act labels used in the corpus.

responding emotional states for positive, and output null vector for negative; (4) verb plus adjective phrase: output vector of adjective phrase.

In the final *Sentence-level Analysis* stage, the overall affect of a sentence and its resulting intensity degrees are estimated. The emotional vector of a simple sentence (or of a clause) is generated from emotional categories and their intensities resulting from phrase-level analysis. AAM also differentiates the strength of the resulting emotion depending on the tense of the sentence and the presence of first person pronouns. For compound sentences, two rules are defined: (1) with coordinate connectors 'and' and 'so': output the vector with the maximum intensity within each corresponding emotional state in the resulting vectors of both clauses; (2) with coordinate connector 'but': the resulting vector of a clause following after the connector is dominant.

For this work, we use the emotion vectors produced by the *Word-level Analysis* and *Sentence-level Analysis* and incorporated them in the learning of the head nod models. In other words, we used the affective labels of each word or the whole sentence. The results of using affective information on word-level and sentence-level are described in the following sections.

3.3. Data Alignment and Feature Selection

One of the main features of our previous work (Nonverbal Behavior Generator [13]) on generating behaviors for virtual humans 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. In addition to this, we investigate if the use of affective information detected from text helps us in learning speaker head nod models.

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 1 lists the different types of dialogue acts used in the corpus. The head movement types annotated in the corpus are: nod, shake, nodshake,

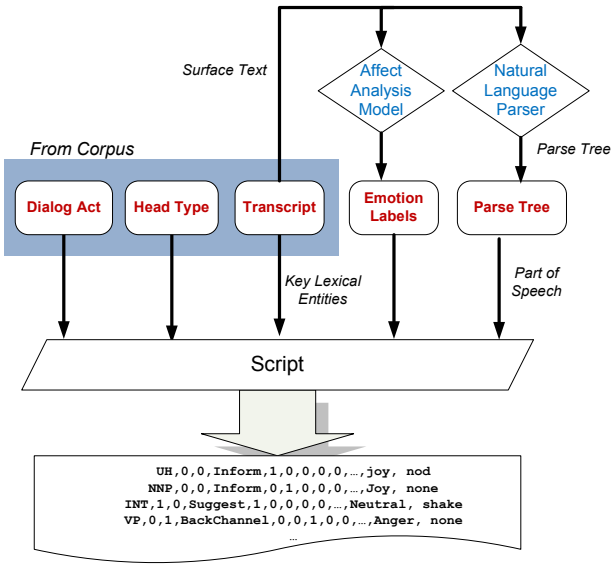

Figure 3. Data Construction Process. From the gesture corpus, speaker transcript, dialog act, and head types are extracted. The transcript is sent to the Affect Analysis Model for emotion labels and a natural language parser for part of speech tags and phrase boundaries. The data set is encoded and transformed into trigrams before being used to train the HMMs.

other, and none. 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 [6]). In addition, we processed the utterances through the Affect Analysis Model to obtain 1) emotion labels for each word (result of word-level analysis stage described in the previous section) and 2) emotion labels for the whole sentence (result of sentence-level analysis stage). Finally, we 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 3 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 features:

- **Part of speech tag** (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
- **Emotion label** (10 cases, if we are using emotion label over whole sentence, words in the same sentence will have

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

Table 2. Features selected for training. The features were selected based on the frequencies of nod-feature co-occurrences. The label 'Remainder' includes everything not falling under other categories.

the same emotion label)
- **Key lexical entities** (keywords that are associated with head nods in NVBG)

Next we reduced the number of features used for learning by eliminating uncorrelated features (i.e. features that do not affect 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. We reduced the number of features by counting the frequency of head nods that occurred with each feature and selected a subset of them. Head nods 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 these results, the final features were selected for training. Table 2 lists the final features.

3.4. Training Process

For this work, we trained the models with three different conditions: 1) using no affective information, 2) using emotion label of each word, 3) using emotion label of the whole sentence. To learn the head nod model, hidden Markov models (HMM) [21] were trained, which is a statistical model widely used for learning patterns where a sequence of observations is given. HMMs have been especially successful in the applications of temporal pattern recognition such as gesture recognition, speech recognition, and part-of-speech tagging [23] [10] [5]. Since learning the patterns of head nods is another case of temporal pattern recognition, we used HMMs to predict head nods. The input for this work is a sequence of feature combinations representing each word.

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 setting then 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 (total of 326 samples) through the 'NOD HMM' and 'NOT_NOD HMM' and classified each sample to have the head movement of whichever model produced a higher probability. We repeated the entire process for the three different conditions described above. The number of states for the final trained model with no emotion, emotion over word, and emotion over sentence were 3, 3, and 2, respectively.

4. Results and Discussion

To measure the performance of our learned model, we computed the accuracy, precision, recall, and F-score of the learned model. Table 3 summarizes the results with the equations used for computing the measurements. Column three replicates some of the result of [14], where no emotion labels were used during learning. Column four and five show the results of incorporating emotion over word and over sentence, respectively. When emotion label for each word was used, the precision rate increased but the recall rate dropped markedly, resulting in a lowered F-score. On the other hand, when emotion label for the whole sentence

Measurement	Without Emotion	Emotion over Word	Emotion over Sentence
Accuracy	.8528	.8589	.8957
Precision	.8249	.9270	.8909
Recall	.8957	.7791	.9018
F-score	.8588	.8467	.8963

Table 3. Measurements for the performance of the learned model.

was used, the accuracy, precision, recall rate all increased, resulting in an improved F-score value.

There are several possible explanations for this. First, it may be that the Affect Analysis Model may perform 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 as described in section 3.2. 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.

5. Conclusions and Future Direction

In this paper we explored the effect of affective information to improve the learning of a probabilistic model for head nods. We extended our previous work [14] which uses the linguistic features of the surface text, including the syntactic/semantic structure of the utterance, and incorporated the emotion labels over words and sentences during learning. Hidden Markov models were trained to predict head nods and we contrasted between the cases of using no emotion labels, emotion over word, and emotion over sentences through accuracy, precision, recall, and F-score values. The results show that using affective information can help predict speaker's nods. However, a simple lookup to determine the emotion for the words can also damage the learning, which emphasizes the need for a deeper analysis to improve the learning.

We can extend this work in several ways. First of all, we can learn the patterns of different head movements other than head nods or other nonverbal behaviors. Additionally, we can learn the different patterns of head movements across different cultures, personalities, or genders.

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, 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.

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

[6] 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.

[7] 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.

[8] 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.

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

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

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

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

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

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

[15] 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.

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

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

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

[19] 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.

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

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

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

[23] 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.